

EP 426: Physics of quantum devices

Course instructor : Kantimay Das Gupta

Lectures are held: Mon (8.30-9.25), Tue (9.30-10.25), Thu (10.30-11.25)

Course Contents:

1. Transport in large and small devices
 - The Boltzmann transport equation (BTE)
 - Using the BTE in simple cases
 - What drives current
2. Electric potential, electrochemical potential, chemical potential, Fermi level...what are these?
 - The band bending concept
 - The self consistent band bending equations (Poisson-Schrodinger etc)
 - Band bending near surfaces, interfaces
 - MOSFET and the 2dimensional electron gas
3. Ballistic transport concept
 - Application to 1D channels
 - Quantum Hall effect
 - Single electron transistor
 -
4. Superconducting devices
 - How a suouerconductor differs from a normal metal
 - Josephson junction
 - SQUID
5. Light emitting devices
 - LED, junction laser
 - VCSEL
 - Quantum Cascade Laser
6. A bag of topics..
 - How are mesoscopic or nano-structures made?
 - Using Quantum devices to define the standard Ohm/Volt/Ampere. How and why?
 - Resistance, capacitor and inductor. Is this set incomplete?
 - Band structure of Si, Ge and GaAs
 - How shape of the Fermi surface is expected to change with electron concentration. Why metals have more complex fermi surfaces?
 - various regions of the EM spectrum and length scales.

Evaluation :

1. Class Quiz : 10-15%
 2. Mid sem : 25-30%
 3. Class Quiz/Term Paper : 20-25%
 4. End sem : 30-40%
-

This material is not a book or the complete course. It gives you some guidelines only. You should always refer to the recommended books to see the standard development and data relating to a chapter.

Contents

| | |
|---|-----------|
| 1 Boltzmann Transport equation : | |
| deviation from equilibrium and current flow | 7 |
| 1.1 A "handwaving" derivation of the equation | 7 |
| 1.2 The semiclassical Boltzmann equation | 8 |
| 1.3 Electric field only | 9 |
| 1.3.1 The temperature dependence of mobility, conductivity | 12 |
| 1.4 Conservation of the phase space volume | 13 |
| 1.5 Electric and magnetic field | 14 |
| 1.6 Moments of the transport equation: Continuity & Drift-diffusion | 17 |
| 1.6.1 Continuity equation | 17 |
| 1.6.2 Drift-diffusion equation | 18 |
| 1.7 Relation of particle current to electrochemical potential and thermal gradients | 19 |
| 1.8 The convective derivative | 21 |
| 2 Charge densities, dopants, junctions and band-bending in semiconductors | 23 |
| 2.1 Carrier concentration and doping | 23 |
| 2.2 A few useful numbers about Si, Ge and GaAs | 25 |
| 2.3 Fermi Level in an intrinsic (undoped) semiconductor | 25 |
| 2.4 Fermi level in a doped semiconductor | 27 |
| 2.4.1 Thermal ionisation (Saha equation) of the dopant system | 30 |
| 2.4.2 General method of solving for the Fermi level | 30 |
| 2.5 Metal-semiconductor and semiconductor-semiconductor junctions | 33 |
| 2.5.1 Situations with no current flow | 33 |
| 2.5.2 How realistic are these calculations? | 36 |
| 2.5.3 When is a contact not a "Schottky" ? | 36 |
| 2.5.4 Situations with varying E_f : what more is needed? | 37 |
| 2.6 The tunnel diode | 42 |
| 2.6.1 The Bipolar Junction Transistor | 45 |
| 3 MOSFETs and GaAs-AlGaAs hetrostructures : accumulation of the 2-dimensional electron gas | 47 |
| 3.1 The MOSFET | 47 |
| 3.1.1 At what voltage will the MOSFET's channel start forming ? | 52 |
| 3.2 The GaAs-AlGaAs heterostructure | 54 |
| 3.2.1 Anderson's rule, alignment of the bands at the heterointerface | 54 |
| 3.2.2 Band structure of the GaAs-AlGaAs heterostructure | 56 |
| 3.3 The envelope function approximation | 58 |
| 3.3.1 A handwaving justification | 59 |
| 3.3.2 Position dependent effective mass | 60 |
| 3.4 Using the envelope function : Fang Howard wavefunction at a heterinterface | 61 |

| | | |
|----------|---|------------|
| 4 | Thomas Fermi screening in the electron gas : Why does remote doping lead to high mobilities? | 63 |
| 4.1 | Polarisation of the lattice | 63 |
| 4.1.1 | Jellium | 63 |
| 4.1.2 | Screened potential | 64 |
| 4.1.3 | The screened Coulomb potential | 65 |
| 4.2 | Screening in the 2-dimensional electron gas | 65 |
| 4.2.1 | Finite thickness with Thomas-Fermi screening | 66 |
| 5 | Quantum Hall effect : Basic physics and the edge state picture | 71 |
| 5.1 | Current flow in a rectangular "Hall Bar", the classical solution | 71 |
| 5.1.1 | $\mathbf{B}=0$ | 71 |
| 5.1.2 | After a magnetic field is applied | 72 |
| 5.1.3 | Current flow directions and the equipotentials | 73 |
| 5.1.4 | What is <i>not</i> correct in the classical solution? | 74 |
| 5.2 | Quantum mechanical solution | 74 |
| 5.2.1 | Degeneracy of each Landau level and the overlap of the wavefunctions | 75 |
| 5.2.2 | Oscillation of E_F as B increases | 75 |
| 5.2.3 | The states near the edge of the sample | 76 |
| 5.2.4 | The "perfect" one-dimensionality of the edge states | 77 |
| 5.2.5 | Disorder broadening of the Landau levels | 80 |
| 5.2.6 | Longitudinal resistance and "Quantum" Hall effect | 80 |
| 6 | Conduction through a constriction: quantisation of conductance | 83 |
| 6.1 | How is such a constriction made? | 83 |
| 6.2 | Calculation of the conductance of a Quantum point contact (QPC) | 85 |
| 7 | Superconductivity | 89 |
| 7.1 | What do we know experimentally | 89 |
| 7.1.1 | It is a distinct thermodynamic phase | 89 |
| 7.1.2 | Evidence for a gap | 91 |
| 7.1.3 | How an attractive e-e interaction becomes possible | 94 |
| 7.2 | Ginzburg Landau "order parameter" theory | 96 |
| 7.2.1 | Another way of writing the current density: flux quantization | 98 |
| 7.3 | The Cooper pair problem | 99 |
| 7.3.1 | Many pairs together | 101 |
| 8 | Josephson junctions | 105 |
| 8.1 | The RCSJ model | 109 |
| 8.1.1 | The titled washboard | 110 |
| 8.1.2 | Two different steady states: role of the initial conditions | 112 |
| 8.1.3 | Introducing the magnetic field | 112 |
| 8.1.4 | Principle of the Superconducting QUantum Interference Device | 114 |
| A | Electromagnetic spectrum | 119 |
| A.1 | Mott transition | 122 |
| B | Many particle systems | 127 |
| B.1 | Determinant and permanent | 127 |
| B.1.1 | Fermions | 127 |
| B.1.2 | Bosons | 128 |
| B.1.3 | An important exercise | 128 |
| B.1.4 | The Hilbert space | 129 |
| B.2 | How do all these help? | 132 |

Chapter 1

Boltzmann Transport equation : deviation from equilibrium and current flow

References:

1. Chapter 12 & 16, *Solid State Physics*, N. W. Ashcroft and N.D. Mermin
 2. page 61-67, *Quantum Heterostructures*, V. V. Mitin, V. A. Kochelap and M. A. Stroscio
 3. A.H. Marshak and K. M. van Vliet, Electrical current in solids with position dependent band structure, *Solid State Electronics*, page 417-427, **21**, 1978
-

Boltzmann Transport equation (BTE) allows us to take a step out of equilibrium thermodynamics. In general the concept of equilibrium implies that there is no net particle flow from one point of a system to another. An equivalent statement is that the electrochemical potential is same throughout the system. Yet, in reality, all electrical devices have currents flowing in them - if it didn't it wouldn't be interesting at all. Often the effects of current flow are also irreversible -like Joule heating. We will see how the presence of electro-magnetic fields, electrochemical potential gradients and thermal gradients drive current. We will do so by trying to calculate the distribution function in a situation slightly away from equilibrium. Our main target is to find expressions for current when "external fields" are present.

1.1 A "handwaving" derivation of the equation

Let's consider the phase space with just two co-ordinates \mathbf{r}, \mathbf{p} . The distribution is represented by a point as shown at time t . What happens to the points in the "volume element" after a little while? Unless they are scattered they change their co-ordinates according to the following rule:

$$\begin{aligned}\mathbf{r}(t + \delta t) &= \mathbf{r}(t) + \frac{\mathbf{p}}{m} \delta t \\ \mathbf{p}(t + \delta t) &= \mathbf{p}(t) + \mathbf{F} \delta t\end{aligned}\tag{1.1}$$

If this was all, then it wouldn't have been interesting at all, because it would mean:

$$f(\mathbf{r}(t) + \frac{\mathbf{p}}{m} \delta t, \mathbf{p}(t) + \mathbf{F} \delta t, t + \delta t) d^3 \mathbf{r}' d^3 \mathbf{p}' = f(\mathbf{r}(t), \mathbf{p}(t), t) d^3 \mathbf{r} d^3 \mathbf{p}\tag{1.2}$$

Now note the following:

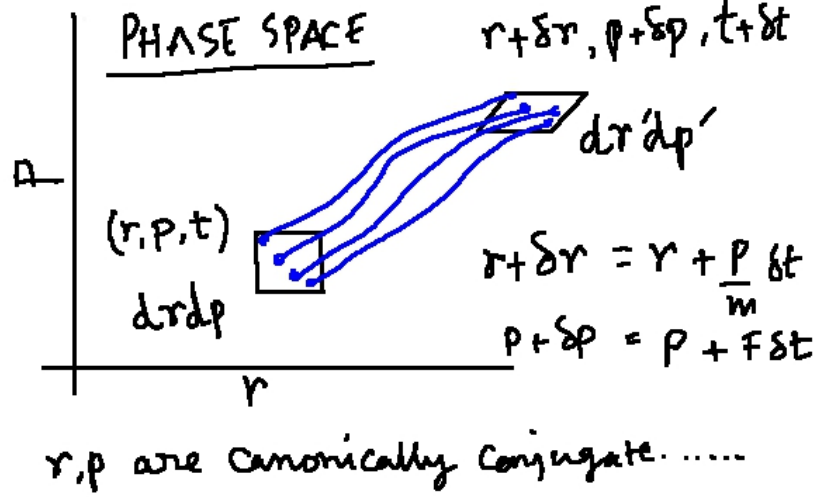


Figure 1.1: Flow of points in phase space.

- The volume element around the point distorts, but preserves its volume. Something that was a square at time t , would become a parallelogram at $t + \delta t$. This would happen if the external force is derivable from a potential and the two co-ordinates that we have chosen are canonically conjugate. We will ignore this technicality at this point. Many text books on statistical physics, treat this point carefully...
- The equality fails to hold, because some of the trajectories get scattered by collisions. Thus the amount by which this equality fails to hold, must be attributed to collisions. This leads us to the following

$$\frac{\mathbf{p}}{m} \cdot \nabla f + \mathbf{F} \cdot \nabla_{\mathbf{p}} f + \frac{\partial f}{\partial t} = \frac{df}{dt} \quad (1.3)$$

Now we take another step, to convert this classical equation into a semiclassical one. We change momentum to wavevector. This indeed means that we are using the concept of phase space (clearly defined momentum and position) in quantum mechanical scenario. An analysis of how far this can give correct results, is non-trivial, but we will give an answer later.

1.2 The semiclassical Boltzmann equation

We denote the equilibrium distribution function by $f^0(\mathbf{r}, \mathbf{k}, t)$. When the distribution function deviates from equilibrium, a "restoring effect" arises in the system, that tries to push the distribution back towards equilibrium. This implies that the collision integral on the right hand side of BTE is assumed to have the form

$$\left. \frac{df}{dt} \right|_{\text{collision}} = -\frac{f - f^0}{\tau} \quad (1.4)$$

Later on we will try to determine τ in terms of the scattering mechanisms in some systems. The best justification of the relaxation time approximation is that it works in many cases!

We thus write the BTE as:

$$\frac{\partial f}{\partial t} + \frac{d\mathbf{r}}{dt} \cdot \nabla_{\mathbf{r}} f + \frac{d\mathbf{k}}{dt} \cdot \nabla_{\mathbf{k}} f = -\frac{f - f^0}{\tau} \quad (1.5)$$

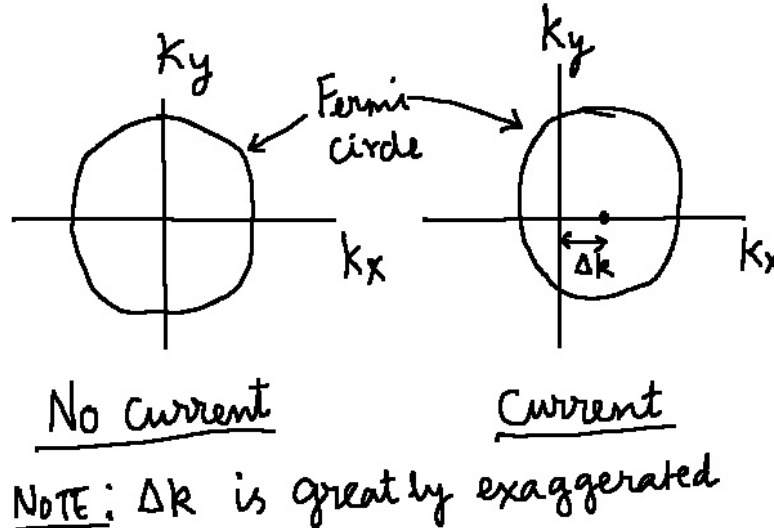


Figure 1.2: Displacement of the Fermi circle results in current flow.

If there are electric and magnetic fields in the system, the "semiclassical" equation of motion would be:

$$\hbar \frac{d\mathbf{k}}{dt} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \quad (1.6)$$

We will also assume that f has no explicit time dependence and that $\nabla_r f = 0$ as well - which in general means that there is no density gradient of particles in the system. This assumption is correct if we are dealing with a piece of copper wire at constant temperature, but not necessarily correct for a semiconductor or a even a piece of metal with a thermal gradient. Throughout the lectures we will assume that the charge of each particle is " q ". For the most common case of electrons in the conduction band we would need to put $q = -|e|$ to get the correct sign of the terms.

1.3 Electric field only

With these assumptions, equation 1.5 in presence of an electric field only reduces to

$$\frac{q}{\hbar} \mathbf{E} \cdot \nabla_k f = -\frac{f - f^0}{\tau} \quad (1.7)$$

Then we make the first order approximation by taking the derivative around the equilibrium value

$$f(\mathbf{k}) = f^0(\mathbf{k}) - \frac{q\tau}{\hbar} \mathbf{E} \cdot \nabla_k f^0 \quad (1.8)$$

$$= f^0\left(\mathbf{k} - \frac{q\tau}{\hbar} \mathbf{E}\right) \quad (1.9)$$

This means that the equilibrium distribution function has retained its functional form but just got shifted by a certain amount. Think of how the graph of a function $f(x)$ would be related to $f(x - a)$. In the figure we have drawn it for a Fermi distribution in 2 dimensions. Note that if the relaxation mechanism is strong then τ would be small. On the other hand if the particle suffers very little scattering then τ would be large and the displacement of the Fermi circle (or sphere) would also be large.

PROBLEM: The free electron density in Copper is $n = 8.5 \times 10^{28} \text{ m}^{-3}$ and near room temperature the relaxation time of most metals is of the order of $10^{-15} - 10^{-14} \text{ sec}$. From this data estimate the

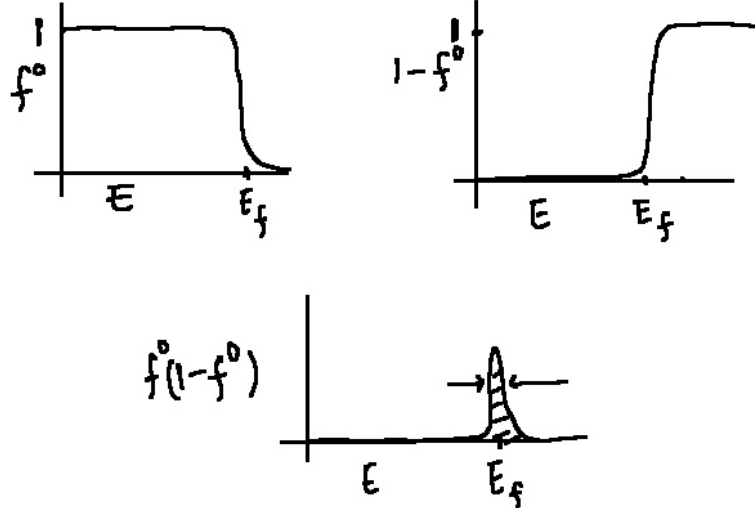


Figure 1.3: How does the product $f^0(1 - f^0)$ behave?

fractional shift of the distribution on the scale of the Fermi wavevector (k_F) for an electric field of 10V/m, (i.e. calculate $\Delta k/k_F$).

Our target is to calculate the current produced by this state:

$$\begin{aligned} \mathbf{j} &= q \sum_{\mathbf{k}} \mathbf{v} \delta f \\ &= \frac{2q}{(2\pi)^3} \int d^3\mathbf{k} \mathbf{v} \delta f \end{aligned} \quad (1.10)$$

To proceed we need to evaluate eqn 1.9 for the case of Fermi distribution.

$$f^0 = \frac{1}{e^{\beta(E-E_f)} + 1} \quad (1.11)$$

$$\begin{aligned} \nabla_{\mathbf{k}} f^0 &= - \left(\frac{1}{e^{\beta(E-E_f)} + 1} \right)^2 e^{\beta(E-E_f)} \nabla_{\mathbf{k}} \beta(E - E_f) \\ &= -\beta f^0(1 - f^0) \nabla_{\mathbf{k}} E \\ &= -\beta f^0(1 - f^0) \hbar \mathbf{v}_{\mathbf{g}} \end{aligned} \quad (1.12)$$

Notice that the Fermi level is not a function of \mathbf{k} . The end result of 1.12 can also be written as :

$$\nabla_{\mathbf{k}} f^0 = \frac{\partial f^0}{\partial E} \hbar \mathbf{v}_{\mathbf{g}} \quad (1.13)$$

Equations 1.12 and 1.13 are important results as these derivatives occur frequently in transport related physics. How does the product $f^0(1 - f^0)$ behave ? Since f^0 drops sharply around E_f , $(1 - f^0)$ must rise sharply around E_f , producing a sharp peak.

PROBLEM : Certain combinations of the Fermi function, occur very frequently in expressions that involve scattering or transitions. It is useful to be familiar with the combination $f^0(1 - f^0)$. Make a rough sketch of how $f^0(1 - f^0)$ would look as a function of energy. How does the area under the curve of $f^0(1 - f^0)$ vary with temperature?

Using eqn 1.12 and eqn 1.9 we get

$$\delta f = q\tau\beta f^0(1 - f^0)\mathbf{E}\cdot\mathbf{v}_g \quad (1.14)$$

Notice that the change occurs only near the Fermi surface. This is the generic reason phenomena like electrical or heat conduction are often referred to as a "Fermi surface property". Now we calculate the current as defined in eqn 1.10

$$\begin{aligned} \mathbf{j} &= \frac{q}{4\pi^3} \int d^3\mathbf{k} \mathbf{v}_g (q\tau\beta f^0(1 - f^0)\mathbf{E}\cdot\mathbf{v}_g) \\ &= nq \left(\frac{q}{4\pi^3n} \int d^3\mathbf{k} \tau\mathbf{v}_g \otimes \mathbf{v}_g \left(-\frac{\partial f^0}{\partial E} \right) \right) \cdot \mathbf{E} \end{aligned} \quad (1.15)$$

Notice that the part within the large brackets is determined by equilibrium properties of the system only. The outer product (\otimes) of two vectors is an object with two indices and can be written out like a matrix. For example

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \quad \text{implies} \quad (1.16)$$

$$C_{ij} = A_i B_j \quad (1.17)$$

We will call the quantity inside the bracket as mobility. But it is often not necessary to evaluate this is full generality. We assume that the dispersion relation is spherically symmetric and evaluate the expression for low temperature. Low temperature implies that the Fermi distribution has a sharp drop near E_f and behaves like a step function at that point. The derivative of a step function is a (Dirac) delta function which would pick out the contribution of the integrand around its peak. So we can write

$$\lim_{T \rightarrow 0} -\frac{\partial f^0}{\partial E} = \delta(E - E_f) \quad (1.18)$$

Let's go through the steps for evaluating the mobility integral:

$$\overleftrightarrow{\mu} = \frac{q}{4\pi^3n} \int d^3\mathbf{k} \tau\mathbf{v}_g \otimes \mathbf{v}_g \left(-\frac{\partial f^0}{\partial E} \right) \quad (1.19)$$

$$= \frac{q}{n} \int dE D(E) \tau\mathbf{v}_g \otimes \mathbf{v}_g \left(-\frac{\partial f^0}{\partial E} \right)$$

$$= \frac{q}{n} \int dE D(E) \tau\mathbf{v}_g \otimes \mathbf{v}_g \delta(E - E_f) \quad \text{as } T \rightarrow 0 \quad (1.20)$$

$$(1.21)$$

Now since $\mathbf{v}_g = \hbar\mathbf{k}/m$, we can write:

$$\mu_{ij} = \frac{q}{n} \int dE D(E) \tau \left(\frac{\hbar}{m} \right)^2 k_i k_j \delta(E - E_f) \quad (1.22)$$

This from works in all dimensions, provided the density n is interpreted correctly. Now μ_{ij} will average to zero if $i \neq j$, due to symmetry. If we fix k_i , we can find corresponding pairs of points at k_j and $-k_j$, which will add up to zero. So we need to calculate only the diagonal terms. Since there is nothing to distinguish the x, y or z directions, all the diagonal components must be equal. This allows us to write:

$$\begin{aligned} \mu_{ii} &= \frac{q}{n} \int dE D(E) \tau \left(\frac{\hbar}{m} \right)^2 \frac{k_x^2 + k_y^2 + k_z^2}{3} \delta(E - E_f) \\ &= \frac{q}{3n} \int dE D(E) \tau \frac{2E}{m} \delta(E - E_f) \end{aligned} \quad (1.23)$$

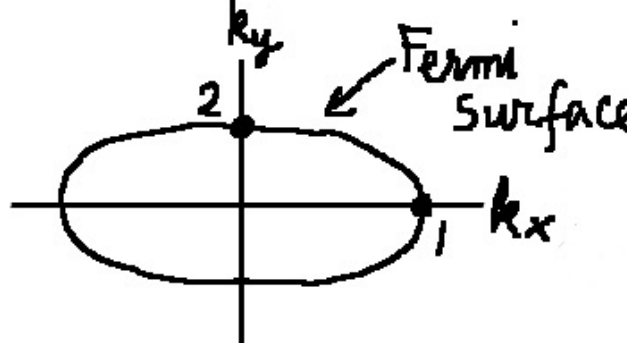


Figure 1.4: Along which direction do you expect the effective mass to be higher, at point 1 (along k_x) or point 2 (along k_y)?

Using the expression for density of states in 3D, eqn 1.23 reduces to:

$$\begin{aligned}\mu_{ii} &= \frac{q}{3\pi^2 m n} \int dE k^3 \tau \delta(E - E_f) \\ &= \frac{q\tau}{m} \quad \text{since } k_F^3 = 3\pi^2 n\end{aligned}\tag{1.24}$$

1.3.1 The temperature dependence of mobility, conductivity

The mobility is a temperature dependent quantity - the T dependence of conductivity for example arises from changes in mobility as well as carrier density of a system. Usually scattering calculations give us the scattering rate ($\tau(E)$) or the collision cross section as a function of E . How do we use this information to calculate $\mu(T)$. Let's consider the diagonal element (say μ_{xx}) from equation 1.15 which relates \mathbf{j} and \mathbf{E}

$$\begin{aligned}\mu_{ij} &= \frac{q \int_0^\infty dE D(E) \tau(E) v_i v_j \left(-\frac{\partial f}{\partial E} \right)}{n} \\ \therefore \mu_{xx} &= \frac{q \int_0^\infty dE D(E) \tau(E) v_x^2 \left(-\frac{\partial f}{\partial E} \right)}{\int_0^\infty dE D(E) f(E)}\end{aligned}$$

Now if we are working in d dimensions, then in general we have

$$\begin{aligned}D(E) &\propto E^{d/2-1} \\ E &\propto m v_x^2 \frac{d}{2}\end{aligned}\tag{1.25}$$

Using these two results and a partial integration of the denominator we get:

$$\begin{aligned}\mu_{xx} &= \frac{2q}{md} \frac{\int_0^\infty dE E^{d/2} \tau(E) \left(-\frac{\partial f}{\partial E}\right)}{\int_0^\infty dE E^{d/2} \left(-\frac{\partial f}{\partial E}\right)} \\ &= \frac{q}{m} \frac{\int_0^\infty dE E^{d/2} \tau(E) \frac{\partial f}{\partial E}}{\int_0^\infty dE E^{d/2} \frac{\partial f}{\partial E}}\end{aligned}$$

Since $\mu = \frac{q\tau}{m}$, we usually write,

$$\langle \tau(T) \rangle = \frac{\int_0^\infty dE E^{d/2} \tau(E) \frac{\partial f}{\partial E}}{\int_0^\infty dE E^{d/2} \frac{\partial f}{\partial E}} \quad (1.26)$$

$\tau(E)$ is often available from scattering calculations and the integral gives the energy range over which we need to average it. The presence of the term $\frac{\partial f}{\partial E}$ ensures that the important part is centred at Fermi energy, the spread of the region increases with increasing temperature.

1.4 Conservation of the phase space volume

We will apply the BTE to a situation where the "forces" will have some velocity dependence, like the Lorentz force. So let's prove that the "volume" will still be conserved. Part of the proof is left as an exercise. We will work with two variables only for simplicity. Consider the points (x, p) and a small area element $\delta x \delta p$ around it as before. What happens to the corner points after time δt ? Both \dot{x} and \dot{p} can be functions of x and p , but we do not write all the functional dependences explicitly. See the following table:

| point | time= t | time= $t + \delta t$ |
|--------------------|---|---|
| $1 \rightarrow 1'$ | $\begin{pmatrix} x \\ p \end{pmatrix}$ | $\begin{pmatrix} x + \dot{x}\delta t \\ p + \dot{p}\delta t \end{pmatrix}$ |
| $2 \rightarrow 2'$ | $\begin{pmatrix} x + \delta x \\ p \end{pmatrix}$ | $\begin{pmatrix} x + \delta x + \left(\dot{x} + \frac{\partial \dot{x}}{\partial x}\delta x\right)\delta t \\ p + \left(\dot{p} + \frac{\partial \dot{p}}{\partial x}\delta x\right)\delta t \end{pmatrix}$ |
| $4 \rightarrow 4'$ | $\begin{pmatrix} x \\ p + \delta p \end{pmatrix}$ | $\begin{pmatrix} x + \left(\dot{x} + \frac{\partial \dot{x}}{\partial p}\delta p\right)\delta t \\ p + \delta p + \left(\dot{p} + \frac{\partial \dot{p}}{\partial p}\delta p\right)\delta t \end{pmatrix}$ |

PROBLEM: Show that the area element $\delta x \delta p$ will become $\delta x' \delta p'$ after time δt where

$$\delta x' \delta p' = \begin{vmatrix} \delta x \left(1 + \frac{\partial \dot{x}}{\partial x} \delta t\right) & \frac{\partial \dot{p}}{\partial x} \delta x \delta t \\ \frac{\partial \dot{x}}{\partial p} \delta p \delta t & \delta p \left(1 + \frac{\partial \dot{p}}{\partial p} \delta t\right) \end{vmatrix} \quad (1.27)$$

Then prove that if x and p are driven by a Hamiltonian such that $\dot{x} = \frac{\partial H}{\partial p}$ and $\dot{p} = -\frac{\partial H}{\partial x}$ then the first order part (in δt) of the expression will be zero. Since equations of motion in a magnetic field can be written in Hamiltonian form as well with canonical momentum defined properly, we can still use the equations.

1.5 Electric and magnetic field

Now let's recall eqn 1.6 and allow a magnetic field. Eqn 1.8 that described the deviation of the distribution function from equilibrium should now read :

$$f(\mathbf{k}) = f^0(\mathbf{k}) - \frac{q\tau}{\hbar} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{k}} f^0 \quad (1.28)$$

Now because the force term has explicit dependence on \mathbf{k} we can no longer write down the solution by inspection, as we did in eqn 1.9. However we now try a solution of the same form, with an unknown vector \mathbf{Z} . Our target is to write \mathbf{Z} as a function of \mathbf{E} and \mathbf{B} , but free of \mathbf{k} and $\mathbf{v}_{\mathbf{g}}$. Thus we want \mathbf{Z} , such that

$$f(\mathbf{k}) = f^0(\mathbf{k} - \frac{q\tau}{\hbar} \mathbf{Z}) \quad (1.29)$$

Hence,

$$\delta f = -\frac{q\tau}{\hbar} \mathbf{Z} \cdot \nabla_{\mathbf{k}} f^0 \quad (1.30)$$

We now use the assumed form (eqn 1.30) with eqns 1.5 and 1.6. This gives:

$$\frac{q}{\hbar} (\mathbf{v} \times \mathbf{B}) \cdot (\nabla_{\mathbf{k}} f^0 + \nabla_{\mathbf{k}} \delta f) + \frac{q}{\hbar} \mathbf{E} \cdot \nabla_{\mathbf{k}} f^0 = -\frac{\delta f}{\tau} \quad (1.31)$$

We already know that $\nabla_{\mathbf{k}} f^0$ points along $\mathbf{v}_{\mathbf{g}}$ and hence the first term in eqn 1.31 gives zero. This leaves us with

$$\frac{q}{\hbar} (\mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{k}} \delta f + \frac{q}{\hbar} \mathbf{E} \cdot \nabla_{\mathbf{k}} f^0 = \frac{q}{\hbar} \mathbf{Z} \cdot \nabla_{\mathbf{k}} f^0 \quad (1.32)$$

Now we need to calculate $\nabla_{\mathbf{k}} \delta f$.

$$\begin{aligned} \nabla_{\mathbf{k}} \delta f &= \nabla_{\mathbf{k}} \frac{q\tau}{\hbar} \mathbf{Z} \cdot \nabla_{\mathbf{k}} f^0 \\ &= \frac{q\tau}{\hbar} \nabla_{\mathbf{k}} (-\beta f^0 (1 - f^0) \mathbf{Z} \cdot \hbar \mathbf{v}_{\mathbf{g}}) \\ &= -\beta q\tau ((1 - f^0)(\mathbf{Z} \cdot \mathbf{v}_{\mathbf{g}}) \nabla_{\mathbf{k}} f^0 + f^0(\mathbf{Z} \cdot \mathbf{v}_{\mathbf{g}}) \nabla_{\mathbf{k}} (1 - f^0) + f^0(1 - f^0) \nabla_{\mathbf{k}} (\mathbf{Z} \cdot \mathbf{v}_{\mathbf{g}})) \end{aligned} \quad (1.33)$$

Once again the first two terms in the RHS of 1.33 will give zero when dotted with $\mathbf{v} \times \mathbf{B}$ as they are $\propto \mathbf{v}_{\mathbf{g}}$. The only term left is

$$\nabla_{\mathbf{k}} \mathbf{Z} \cdot \mathbf{v}_{\mathbf{g}} = \nabla_{\mathbf{k}} \mathbf{Z} \cdot \frac{\hbar \mathbf{k} - q \mathbf{A}}{m} = \frac{\hbar}{m} \mathbf{Z} \quad (1.34)$$

In eqn 1.34, \mathbf{A} denotes the vector potential of the magnetic field, \mathbf{v}_g is related to the canonical momentum in presence of a magnetic field in the usual way. Combining eqns 1.33 and 1.34 we can write:

$$(\mathbf{v} \times \mathbf{B}) \cdot \nabla_k \delta f = -\beta q \tau f^0 (1 - f^0) (\mathbf{v} \times \mathbf{B}) \cdot \frac{\hbar}{m} \mathbf{Z} \quad (1.35)$$

So eqn 1.32 now simplifies to:

$$\begin{aligned} & -\frac{\hbar}{m} \beta q \tau f^0 (1 - f^0) (\mathbf{v} \times \mathbf{B}) \cdot \mathbf{Z} + (\mathbf{E} - \mathbf{Z}) \cdot \nabla_k f^0 = 0 \\ \therefore & -\frac{\hbar}{m} \beta q \tau f^0 (1 - f^0) (\mathbf{v} \times \mathbf{B}) \cdot \mathbf{Z} + (\mathbf{E} - \mathbf{Z}) \beta f^0 (1 - f^0) \hbar \mathbf{v}_g = 0 \\ \therefore & \frac{q \tau}{m} (\mathbf{v}_g \times \mathbf{B}) \cdot \mathbf{Z} + (\mathbf{E} - \mathbf{Z}) \cdot \mathbf{v}_g = 0 \\ \therefore & \frac{q \tau}{m} (\mathbf{B} \times \mathbf{Z}) \cdot \mathbf{v}_g + (\mathbf{E} - \mathbf{Z}) \cdot \mathbf{v}_g = 0 \\ \therefore & \mathbf{E} = \mathbf{Z} - \frac{q \tau}{m} \mathbf{B} \times \mathbf{Z} \end{aligned} \quad (1.36)$$

We call \mathbf{Z} as the Hall vector. When both \mathbf{E} and \mathbf{B} fields are present, this quantity in some way, "replaces" the electric field in the transport equation. But we still need to express \mathbf{Z} explicitly in terms of \mathbf{E} and \mathbf{B} , with $\mu = q\tau/m$. The proof is left as an exercise.

$$\mathbf{Z} = \frac{\mathbf{E} + \mu \mathbf{B} \times \mathbf{E} + \mu^2 (\mathbf{B} \cdot \mathbf{E}) \mathbf{B}}{1 + \mu^2 B^2} \quad (1.37)$$

PROBLEM: If $\mathbf{E} = \mathbf{Z} - \mathbf{A} \times \mathbf{Z}$, then show that

$$\mathbf{Z} = \frac{\mathbf{E} + \mathbf{A} \times \mathbf{E} + (\mathbf{A} \cdot \mathbf{E}) \mathbf{A}}{1 + A^2}$$

Hint : Try $\mathbf{A} \times \mathbf{E}$ and $\mathbf{A} \cdot \mathbf{E}$

Recall that the relation current with electric field was reduced to a simple (Drude) form for simple parabolic $E(\mathbf{k})$ dispersion. Following this we then write the expression for current in presence of a magnetic field by replacing \mathbf{E} by \mathbf{Z} :

$$\mathbf{j} = nq\mu \mathbf{Z} = \sigma_0 \mathbf{Z} \quad (1.38)$$

A very general expression with arbitrary \mathbf{E} and \mathbf{B} can be written, but is not very useful. Rather, we consider a situation where the magnetic field points along $\hat{\mathbf{z}}$, and the electric field is in the $x-y$ plane. So we have :

$$\begin{aligned} \mathbf{E} &= E_x \hat{\mathbf{x}} + E_y \hat{\mathbf{y}} \\ \mathbf{B} &= B_0 \hat{\mathbf{z}} \end{aligned} \quad (1.39)$$

and hence:

$$\begin{aligned} Z_x &= \frac{E_x - \mu B_0 E_y}{1 + \mu^2 B_0^2} \\ Z_y &= \frac{E_y + \mu B_0 E_x}{1 + \mu^2 B_0^2} \end{aligned}$$

Eqn 1.38 then can be written out in 2×2 matrix form as :

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = \frac{\sigma_0}{1 + \mu^2 B_0^2} \begin{pmatrix} 1 & -\mu B_0 \\ \mu B_0 & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad (1.40)$$

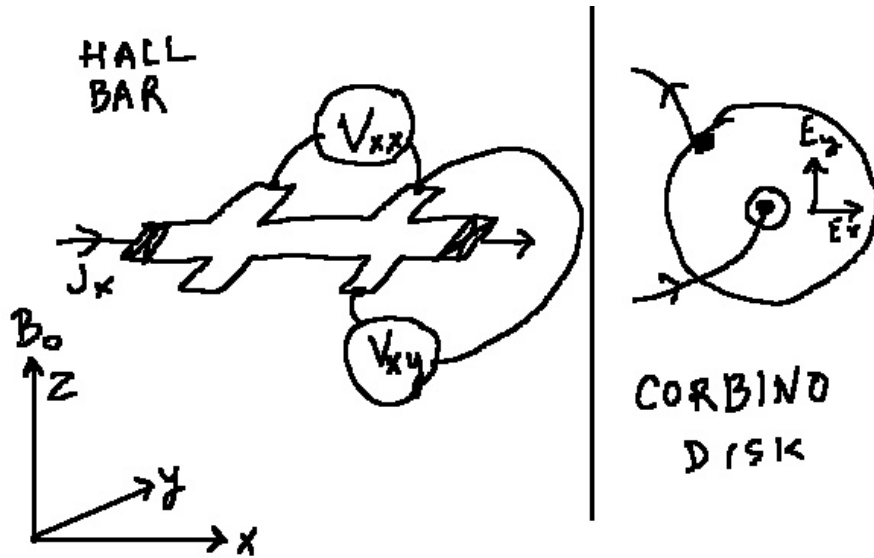


Figure 1.5: Two device geometries commonly used in experiments with 2-dimensional systems

Eqn 1.40 can be inverted to give the resistivity matrix such that:

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} \rho_0 & \frac{B_0}{nq} \\ -\frac{B_0}{nq} & \rho_0 \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} \quad (1.41)$$

where we have written ρ_0 for $1/\sigma_0$.

How do we relate 1.41 to experimental situations? Consider a rectangular block in the xy plane, with the current injecting contacts placed as shown. Sufficiently away from the contacts, the current component j_y must vanish, because there are no current sourcing/withdrawing contacts on the long sides. This allows us to interpret the ratio E_x/j_x as the longitudinal voltage drop and E_y/j_x as the Hall (transverse) voltage. The off-diagonal terms are linear in B and offers the most common way of measuring the electron density in a 2-dimensional system.

PROBLEM: Consider the "Corbino-disk" geometry shown in figure. Current flow is between the inner (central) contact and the outer (circumferential) contact. Show by symmetry arguments that one of the components of the electric field (E_y in figure) must be zero. Can you roughly sketch the current flow paths from the center to the circumference?

It is important to understand that resistance or conductance can no longer be specified by a single number in presence of a magnetic field. They must be understood in a matrix sense. In fact by inverting the resistivity matrix you can easily show that in a magnetic field both σ_{xx} and ρ_{xx} can be simultaneously zero, which appears counter-intuitive at first glance - but there is no contradiction in it.

PROBLEM: Invert the matrix in the equation 5.11. Call this the conductivity matrix whose elements are σ_{ij} . What will be the value of ρ_{xx} if $\sigma_{xx} = 0$, when no magnetic field is present? How would your answer be modified when a finite strong magnetic field is present?

1.6 Moments of the transport equation: Continuity & Drift-diffusion

Taking the moments of a differential equation means multiplying both sides of the equation with some function and integrating over all states/ space. How does that help? The integration "removes" some variable and results in a simpler looking equation. Of course the "simpler" equation is no longer as detailed or informative as the original one - but sometimes we may need focus on a broad feature while removing some details. The BTE refers to the distribution function which is not always possible (or necessary) to know. We show cases where focussing on quantities averaged over the distribution $f(\mathbf{r}, \mathbf{k})$ is immensely useful. The simple 1D version reads:

1.6.1 Continuity equation

This is an expected result of course. However it is useful to show the process. The distribution function $f(\mathbf{r}, \mathbf{k}, t)$ is written as f for simplicity:

$$\frac{\partial f}{\partial t} + \frac{q}{\hbar} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{k}} f + \mathbf{v} \cdot \nabla_{\mathbf{r}} f = \left. \frac{df}{dt} \right|_{\text{collision}} \quad (1.42)$$

We integrate/sum over all \mathbf{k} states. The first term in LHS gives

$$\int \frac{d^3 \mathbf{k}}{(2\pi)^3} \frac{\partial f}{\partial t} = \frac{\partial n(\mathbf{r})}{\partial t} \quad (1.43)$$

where $n(\mathbf{r})$ is the conventional particle density at \mathbf{r} .

To evaluate the second term in LHS use the following vector identity. f is a scalar and \mathbf{A} is a vector.

$$\mathbf{A} \cdot \nabla f = \nabla \cdot f \mathbf{A} - f \nabla \cdot \mathbf{A} \quad (1.44)$$

This implies:

$$(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{k}} f = \nabla_{\mathbf{k}} \cdot f (\mathbf{E} + \mathbf{v} \times \mathbf{B}) - f \nabla_{\mathbf{k}} \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \quad (1.45)$$

The first term can be converted to a surface integral. Because the surface will grow as k^2 , the fields are finite and the Maxwell-Boltzmann and Fermi distributions go to zero as $\sim e^{-k^2}$, this integral will vanish. The next term is also zero, provided we interpret \mathbf{v} as the group velocity. The steps are left as an exercise. Here \mathcal{E} denotes energy.

$$\begin{aligned} \nabla_{\mathbf{k}} \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B}) &= \frac{\partial E_i}{\partial k_i} + \varepsilon_{ijk} \frac{\partial}{\partial k_i} \frac{\partial \mathcal{E}}{\partial k_j} B_k \\ &= 0 + \varepsilon_{ijk} \frac{\partial^2 \mathcal{E}}{\partial k_i \partial k_j} B_k \\ &= 0 + 0 \end{aligned} \quad (1.46)$$

Now, the third term in LHS of 1.42 is

$$\begin{aligned} \int \frac{d^3 \mathbf{k}}{(2\pi)^3} \mathbf{v} \cdot \nabla_{\mathbf{r}} f &= \nabla_{\mathbf{r}} \cdot \int \frac{d^3 \mathbf{k}}{(2\pi)^3} \mathbf{v} f \\ &= \nabla_{\mathbf{r}} \cdot n(\mathbf{r}) \langle \mathbf{v} \rangle \end{aligned} \quad (1.47)$$

The RHS of 1.42 must give zero when integrated over all k -space because the particles which are scattered out of a certain volume must be appearing in some other volume.

Adding 1.43, 1.46, 1.47 gives

$$\frac{\partial n(\mathbf{r})}{\partial t} + \nabla_{\mathbf{r}} \cdot n(\mathbf{r}) \langle \mathbf{v} \rangle = 0 \quad (1.48)$$

1.6.2 Drift-diffusion equation

We multiply both sides of the BTE by velocity (or momentum) and integrating over all states. Obviously the RHS will give the general expression for current. The LHS will be formed of two or three terms with distinct physical meaning. This is very extensively used in describing electronic transport in metals and semiconductors. There are important assumptions which go into it and one needs to be aware of those! The calculation is somewhat long, so we first do a simplified version in 1D with an electric field only. Then we will show how to generalise it to 3D with both electric and magnetic field.

$$\frac{\partial f}{\partial t} + \frac{q}{\hbar} E \frac{\partial f}{\partial k} + v \frac{\partial f}{\partial x} = -\frac{f - f^0}{\tau} \quad (1.49)$$

multiply by v and integrate over all k . The first term in LHS of 1.49 gives:

$$\begin{aligned} \int \frac{dk}{2\pi} v \frac{\partial f}{\partial t} &= \int \frac{dk}{2\pi} \frac{\partial}{\partial t} f v \\ &= \frac{\partial}{\partial t} n \langle v \rangle \end{aligned} \quad (1.50)$$

The second term in LHS of 1.49 gives with $v = \frac{\hbar k}{m}$:

$$\begin{aligned} \int \frac{dk}{2\pi} v \frac{\partial f}{\partial k} &= \int \frac{dk}{2\pi} \left(\frac{\partial}{\partial k} f v - f \frac{\partial v}{\partial k} \right) \\ &= f v \Big|_{-\infty}^{\infty} - \frac{\hbar}{m} n \end{aligned} \quad (1.51)$$

The third term in LHS of 1.49 gives:

$$\begin{aligned} \int \frac{dk}{2\pi} v^2 \frac{\partial f}{\partial x} &= \frac{\partial}{\partial x} \int \frac{dk}{2\pi} v^2 f \\ &= \frac{\partial}{\partial x} n(x) \langle v^2 \rangle \\ &= \frac{\partial}{\partial x} n(x) \left\langle \frac{2E}{m} \right\rangle \\ &= \frac{kT}{m} \frac{\partial n(x)}{\partial x} \end{aligned} \quad (1.52)$$

Notice the use of thermal average kinetic energy in the last step. This will ultimately lead to a relation between mobility and diffusion constant.

The RHS term :

$$\begin{aligned} - \int \frac{dk}{2\pi} v \frac{f - f^0}{\tau} &= \frac{1}{\tau} \int \frac{dk}{2\pi} f v \\ &= \frac{1}{\tau} n \langle v \rangle \end{aligned} \quad (1.53)$$

Now we can put the last four result together, multiply with τ allover and write $J = nq\langle v \rangle$ for the electric current:

$$\begin{aligned} \tau \frac{\partial}{\partial t} n \langle v \rangle + n \langle v \rangle + \underbrace{\frac{q\tau}{\hbar} E \left(-\frac{\hbar}{m} n \right)}_{\text{drift: } \mu = \frac{q\tau}{m}} + \underbrace{\tau \frac{kT}{m} \frac{\partial n(x)}{\partial x}}_{\text{diffusion: } \propto -D \frac{\partial n(x)}{\partial x}} &= 0 \end{aligned} \quad (1.54)$$

Notice that the ratio of the drift mobility to diffusion constant is $\frac{kT}{q}$, called the Einstein relation. This is correct for a classical distribution only. Notice how \hbar has disappeared, another indication that the result is essentially classical. The relation between drift and diffusion components would be different if full Fermi-Dirac distribution used. However at room temperatures in most devices this holds very well for motion of electrons/holes in a band.

Drift diffusion in 3D

PROBLEM: Now let us remove the simplifying assumptions and take the moment of the BTE after multiplying with \mathbf{v} . We need to work with

$$\underbrace{\int \frac{d^3\mathbf{k}}{(2\pi)^3} \mathbf{v} \frac{\partial f}{\partial t}}_{\text{LHS 1}} + \underbrace{\frac{q}{\hbar} \int \frac{d^3\mathbf{k}}{(2\pi)^3} \mathbf{v} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_{\mathbf{k}} f}_{\text{LHS 2}} + \underbrace{\int \frac{d^3\mathbf{k}}{(2\pi)^3} \mathbf{v} \mathbf{v} \cdot \nabla_{\mathbf{r}} f}_{\text{LHS 3}} = - \int \frac{d^3\mathbf{k}}{(2\pi)^3} \mathbf{v} \frac{f - f^0}{\tau} \quad (1.55)$$

Prove the following results. The algebra can be done using the ϵ - δ notation for handling indices of vectors and tensors. The relation between velocity and the wavevector is $m\mathbf{v} = \hbar\mathbf{k} - q\mathbf{A}$

1. LHS 1 : This gives

$$\frac{\partial}{\partial t} n(\mathbf{r}) \langle \mathbf{v} \rangle$$

2. LHS 2 : Notice the occurrence of the *averaged velocity* in the Lorentz term. The calculation is somewhat non-trivial. Do it carefully! You should get

$$-\frac{q}{m} n(\mathbf{r}) (\mathbf{E} + \langle \mathbf{v} \rangle \times \mathbf{B})$$

3. LHS 3 : The diffusion term requires averaging over the distribution. You should get

$$\nabla_{\mathbf{r}} \cdot n(\mathbf{r}) \langle v_i v_j \rangle$$

For Maxwell-Boltzmann distribution $\langle v_i v_j \rangle = \frac{kT}{m} \delta_{ij}$

4. RHS : This gives the current term

$$-n(\mathbf{r}) \frac{\langle \mathbf{v} \rangle}{\tau}$$

Adding all the results will give the drift diffusion relation.

1.7 Relation of particle current to electrochemical potential and thermal gradients

The preceding sections may give you the impression that current (particle flow) must be associated with electric field or a magnetic field. This is not true. There are very striking instances where there is a strong electric field but no charge flow. Also at the end of this section we will be able to answer the question - what does a voltmeter measure? We have more or less got accustomed to the idea that connecting a voltmeter between two points would measure the line integral of the electric field ("potential difference") between the points. While this is indeed true in many circumstances there are striking situations where it is not true. For example if you look near the surface of a metal or a semiconductor or a p-n junction you will find very strong electric field at equilibrium. But a voltmeter connected between the surface of a metal and somewhere inside would give zero. So would a voltmeter connected across a p-n junction in equilibrium. The purpose of this section is to show that particle flow is ultimately related to the gradients of the "electrochemical potential" (Fermi level) and thermal gradients (if any) in the system. It is an important conceptual point for treating electrons moving in conduction band of a semiconductor that may have spatial variation due to changes in composition

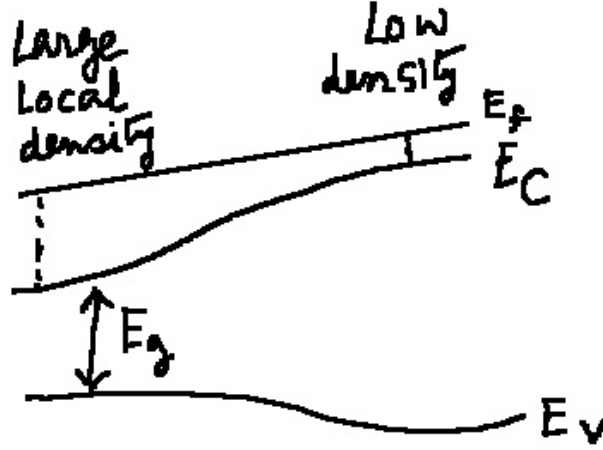


Figure 1.6: The band gap (E_g) may vary due to change in composition of the material. The electrochemical potential (E_f) may vary due to externally applied potentials, like connecting a battery to certain points of the device. The bottom of the conduction band is a profile of the scalar potential - whose gradient gives the electric field at that point.

or effects of accumulated charges or external gates.

Recall the transport equation (eqn 1.5), which we wrote as:

$$\frac{\partial f}{\partial t} + \frac{d\mathbf{r}}{dt} \cdot \nabla_{\mathbf{r}} f + \frac{d\mathbf{k}}{dt} \cdot \nabla_{\mathbf{k}} f = -\frac{f - f^0}{\tau}$$

Now so far we have assumed that $\nabla_{\mathbf{r}} f = 0$, and interpreted $d\mathbf{k}/dt$ as the force on the particle. The force was easily identified as arising from external electric and magnetic fields. When a particle moves in a position dependent band structure, the identification of a "force" on it is not easy - in fact even conceptually it is not absolutely clear (see the paper by Marshak and van Vliet). However this is a commonly encountered situation and an approximation based on Hamilton's equations of motion does work to a good extent.

Given a Hamiltonian $H(\mathbf{r}, \mathbf{p})$ The Hamilton's (classical) equation of motion gives:

$$\frac{d\mathbf{r}}{dt} = \nabla_{\mathbf{p}} H \quad (1.56)$$

$$\frac{d\mathbf{p}}{dt} = -\nabla_{\mathbf{r}} H \quad (1.57)$$

In a "semiclassical" sense we assume that the expectation value of the Hamiltonian can be position-dependent, then we can assign the following meaning to the derivatives. (Note that in a quantum mechanical sense, this can of course be questioned. The reason is that the energy eigenvalues of a Hamiltonian belong to the entire system and are never position dependent, as we are implying here.)

$$\frac{d\mathbf{r}}{dt} = \nabla_{\mathbf{p}} H \approx \mathbf{v}_{\mathbf{g}} \quad (1.58)$$

$$\frac{d\mathbf{p}}{dt} = -\nabla_{\mathbf{r}} H \approx -\nabla_{\mathbf{r}} E \quad (1.59)$$

Using eqns 1.58 and 1.59, the deviation function can be written as:

$$\delta f = -\tau \left(\mathbf{v}_g \cdot \nabla_r f^0 - \frac{1}{\hbar} \nabla_r E \cdot \nabla_k f^0 \right) \quad (1.60)$$

The expression for current is then

$$\mathbf{j} = -\tau \frac{2q}{(2\pi)^3} \int d^3\mathbf{k} \mathbf{v}_g \left(\mathbf{v}_g \cdot \nabla_r f^0 - \frac{1}{\hbar} \nabla_r E \cdot \nabla_k f^0 \right) \quad (1.61)$$

We have calculated the \mathbf{k} -gradient of f^0 before (eqns 1.12 & 1.13), similarly we find that,

$$\nabla_r f^0 = \frac{\partial f^0}{\partial E} \left(\nabla_r (E - E_f) - \frac{E - E_f}{T} \nabla_r T \right) \quad (1.62)$$

Note that E , E_f and T can all have position dependence - but only E has \mathbf{k} dependence. Using eqns 1.13 & 1.62, we can write the current as

$$\begin{aligned} \mathbf{j} &= -\frac{2q}{(2\pi)^3} \int d^3\mathbf{k} \mathbf{v}_g \left(\mathbf{v}_g \cdot \left(\nabla_r E - \nabla_r E_f - \frac{E - E_f}{T} \nabla_r T \right) - \nabla_r E \cdot \mathbf{v}_g \right) \tau \frac{\partial f^0}{\partial E} \\ &= \frac{2q}{(2\pi)^3} \int d^3\mathbf{k} \mathbf{v}_g \left(\mathbf{v}_g \cdot \nabla_r E_f + \frac{E - E_f}{T} \nabla_r T \right) \tau \frac{\partial f^0}{\partial E} \\ &= n \frac{q}{4\pi^3 n} \int d^3\mathbf{k} \left(\mathbf{v}_g \otimes \mathbf{v}_g \tau \frac{\partial f^0}{\partial E} \right) \cdot \nabla_r E_f \\ &\quad + \frac{q}{4\pi^3} \int d^3\mathbf{k} \left(\mathbf{v}_g \otimes \mathbf{v}_g (E - E_f) \tau \frac{\partial f^0}{\partial E} \right) \cdot \frac{\nabla_r T}{T} \end{aligned} \quad (1.63)$$

Note the cancelation of the first and the last terms in step 1. We have already defined the expression in the first part before as mobility (eqns 1.15 & 1.19). The next part gives the particle flow (and hence electric current) as result of a thermal gradient. This thermoelectric coefficient is conventionally denoted with L . With the definitions

$$\overleftrightarrow{\mu} = \frac{q}{4\pi^3 n} \int d^3\mathbf{k} \mathbf{v}_g \otimes \mathbf{v}_g \tau \left(-\frac{\partial f^0}{\partial E} \right) \quad (1.64)$$

$$\overleftrightarrow{L} = \frac{q}{4\pi^3} \int d^3\mathbf{k} \mathbf{v}_g \otimes \mathbf{v}_g (E - E_f) \tau \left(-\frac{\partial f^0}{\partial E} \right) \quad (1.65)$$

we can write

$$\mathbf{j} = n \overleftrightarrow{\mu} \cdot (-\nabla_r E_f) + \frac{\overleftrightarrow{L}}{T} \cdot (-\nabla_r T) \quad (1.66)$$

To see the connection with the simple case of an electric field only we need to put $E_f = qV$, where V is the scalar potential. In that case $\nabla_r E_f = -q\mathbf{E}$. This would lead to the results of section 1. In general if the electrochemical potential and the bottom of the conduction band are parallel, one would recover the proportionality of current and the electric field.

1.8 The convective derivative

The combination of derivatives that occurs in the BTE, had the form

$$\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) \text{something} \quad (1.67)$$

This combination of derivatives will occur whenever we try to follow the rate of change of something while "moving with the flow". This situation is very common in hydrodynamics. We will see that the basic equation of hydrodynamics naturally contains this combination. Think of following a small

incompressible element of the fluid along a streamline. The volume element which was at \mathbf{r} at time t would be found at $\mathbf{r} + \delta\mathbf{r}$ after a time δt .

$$\delta\mathbf{r} = \mathbf{v}\delta t \quad (1.68)$$

Then the change of velocity of the volume element is given by: $\mathbf{r} + \delta\mathbf{r}$ is given by:

$$\mathbf{v}(x + v_x\delta t, y + v_y\delta t, z + v_z\delta t) - \mathbf{v}(x, y, z) \quad (1.69)$$

This change in velocity must be accounted for by the forces acting on this volume element. We will assume for simplicity that the liquid is incompressible. The forces acting on a fluid arise from three sources - pressure, external forces like gravity, viscous drag due to the neighbouring volume elements

$$\rho \left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) \mathbf{v} = \mathbf{f}_{pressure} + \mathbf{f}_{gravity} + \mathbf{f}_{viscosity} \quad (1.70)$$

It is easy to show that

$$\begin{aligned} \mathbf{f}_{pressure} &= -\nabla P \\ \mathbf{f}_{gravity} &= -\nabla \rho g z \end{aligned} \quad (1.71)$$

The viscous drag requires a bit more work. The i^{th} component of the force on an infinitesimal surface $d\mathbf{s}$ immersed in a fluid is given by

$$f_i = S_{ij} ds_j \quad (1.72)$$

where S_{ij} are the elements of the stress tensor. For an incompressible fluid of viscosity η it is:

$$S_{ij} = \eta \left(\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j} \right) \quad (1.73)$$

To calculate the total force on a small volume element $dxdydz$ we need to add the forces acting on the six surfaces of the cube, taking opposite faces in pairs. The calculation is straightforward and the final result is that

$$\mathbf{f}_{viscous} = \eta \nabla^2 \mathbf{v} \quad (1.74)$$

Putting the expression for all the forces together we get the equation

$$\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) \mathbf{v} = -\frac{\nabla P}{\rho} - \nabla g z + \eta \nabla^2 \mathbf{v} \quad (1.75)$$

This is the well known Navier-Stokes equation for incompressible flow. Indeed the simplifying assumption of incompressibility is not entirely correct. No elastic waves like sound would have propagated in the liquid if this was so. For a more detailed discussion see chapter 41 of Feynman lectures in physics (vol 2).

Chapter 2

Charge densities, dopants, junctions and band-bending in semiconductors

References:

1. Chapter 5 (4th edition), *Solid State Electronic Devices*, B. G. Streetman
 2. Chapter 3 *Semiconductor Physics*, K Seeger
 3. Greg Snider's homepage has the tool used to calculate band structures.
See <www.nd.edu/~gsnider>
-

Our target is to answer the following questions :

- How many carriers are there in the bands?
- How many dopants ionize? Where is the Fermi level? What is the driving equation?
- How can we qualitatively sketch the bending of the bands near a surface, metal-semiconductor contact, p-n junctions and heterointerfaces?
- Finally, what (self consistent) equations relate the charge densities and band profiles?

2.1 Carrier concentration and doping

At $T = 0$ in a pure semiconductor, the conduction band is empty and the valence band is full. A completely full or a completely empty band cannot carry current. We will see soon that under these circumstances the Fermi energy lies in the gap between the valence and conduction band. The density of states at the Fermi level is zero. The semiconductor is an insulator at this point.

In reality there is no qualitative distinction between semiconductors and insulators. The distinction is that the bandgap of an insulator is large - *e.g.* Silicon oxide has $E_g \sim 9\text{eV}$, Diamond has $E_g \sim 5\text{eV}$ and so on. The bandgap of typical semiconductors is in the range of nearly zero to 3-4 eV. At very high temperatures, if an insulator hasn't already melted, it will act as a semiconductor.

Carriers in a semiconductor's bands come from two sources:

1. Thermally excited electrons in conduction band and the corresponding vacancies left behind in the valence band.
2. Some *suitable* foreign atoms called dopants which can put some electrons in CB or capture some electrons from VB. Sometimes crystal defects can also play the role of foreign atoms.

Consider a group V atom like Phosphorous replacing an atom of group IV Silicon in the lattice. It has one more electron compared to Si. We keep aside the question about how to get the P atom to replace the Si for the time being - but that is not a trivial question. A "dopant" will not work as a dopant if it does not sit in the right place. It is possible for a P atom to somehow go in as an "interstitial", that will not work. Also the same atom may act as an acceptor or a donor in some cases. For example if Si is incorporated in GaAs lattice, replacing a Ga atom, it will act as a donor. If it replaces an As atom it will act as an acceptor. You can figure out the reason.

The crucial fact is that the binding energy of that remaining electron becomes very low. We give a very simplified model - usually called the "hydrogenic impurity model". We assume that the outermost electron in P behaves as if it is tied to a hypothetical nucleus - that is the P^+ ion core. The binding energy and Bohr radius of an H atom (1s state) is

$$E = -\frac{me^4}{8\varepsilon_0^2\hbar^2} \quad (2.1)$$

$$a_B = 4\pi\varepsilon_0\frac{\hbar^2}{me^2} \quad (2.2)$$

Now we make two crucial claims. Inside the "medium" the free electron mass would be modified such that $m \rightarrow m_{eff}$ and $\varepsilon_0 \rightarrow \varepsilon_0\varepsilon_r$. Typically $\varepsilon_r \sim 10 - 15$ for most semiconductor lattices and $m_{eff} \sim 0.1m$. That means the binding energy would reduce by a factor of ~ 1000 and the Bohr radius would increase by a factor of about ~ 100 . So instead of $E = 13.6\text{eV}$ the binding energy will be a few 1-10 meV, the Bohr radius will increase from 0.5\AA to $\sim 50\text{\AA}$. This means that the electron will be exploring something of the order of a $10 \times 10 \times 10$ lattice units. This in retrospect justifies the use of the lattice dielectric which is a quantity meaningful only if averaged over sum volume of the lattice. Also the fact that the electron gets spread over a large area, means that replacing the free electron mass with the band effective mass can be justified. If the binding energy drops to a few meV, it is clear that at room temperature ($k_B T = 25\text{meV}$) these can be almost fully ionised. The order of magnitude of these numbers ensure that semiconductors can be useful at room temperature.

How can we make the arguments better for using the band effective mass? For a direct gap semiconductor we proceed as follows: We need to treat the extra potential introduced by the impurity atom as a "perturbation" and then solve for the wave function.

$$H = H_0 - \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{e^2}{r} \quad (2.3)$$

where $H_0 = T + V = \left(-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right)$ is the KE + lattice periodic part. m is the free electron mass. The relevant solution to this must be formed out of the eigenfunctions of H_0 , the Bloch functions (mostly from near the bottom of the band). So we write ($C_{n\mathbf{k}}$ are the linear coefficients)

$$\Psi = \sum_{n,\mathbf{k}} C_{n\mathbf{k}} u_{n,\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} \quad (2.4)$$

First justify that the following approximation should work.

$$\Psi \approx u_{n,0} \underbrace{\sum_{\mathbf{k}} C_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}}_{F(\mathbf{r})} \quad (2.5)$$

For $F(\mathbf{r})$, it leads to the effective equation -

$$\left(-\frac{\hbar^2}{2m_{eff}}\nabla^2 + \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{e^2}{r}\right) F(\mathbf{r}) = EF(\mathbf{r}) \quad (2.6)$$

Work out how the m_{eff} appeared here. We have done very similar things before - so the steps are left as an exercise. The justification of a similar model for indirect (and multiple valley) semiconductors like Si is more involved (this was given by Luttinger). The summary of the result is that for direct gap single valley (GaAs) the shallow donor level cluster around one number. See the figure 2.1. For indirect gap Si, the clustering does not happen so well. We will in general take this numbers as experimentally determined parameters.

2.2 A few useful numbers about Si, Ge and GaAs

Table 2.1: List of commonly used parameters of Silicon, Germanium and Gallium Arsenide

| | Silicon | Germanium | Gallium Arsenide |
|---|-----------------------|-----------------------|----------------------|
| Atoms cm^{-3} | 5.0×10^{22} | 4.4×10^{22} | 4.4×10^{22} |
| Crystal structure | Diamond | Diamond | Zincblende |
| Density (gm cm^{-3}) | 2.33 | 5.33 | 5.32 |
| Dielectric constant | 11.9 | 16 | 13.1 |
| Electron affinity (eV) | 4.1 | 4.0 | 4.1 |
| Effective density of states in conduction band at 300K: $N_c(\text{cm}^{-3})$ | 2.8×10^{19} | 1.04×10^{19} | 4.7×10^{17} |
| Effective density of states in valence band at 300K: $N_v(\text{cm}^{-3})$ | 1.04×10^{19} | 6.0×10^{18} | 7.0×10^{18} |
| Band gap at 300K (eV) | 1.12 | 0.66 | 1.42 |
| Intrinsic carrier concentration at 300K: $n(\text{cm}^{-3})$ | 1.5×10^{10} | 2.4×10^{13} | 1.8×10^6 |
| electron effective mass : in units of m_0 , the free electron mass | 0.98, 0.19 | 1.64, 0.082 | 0.067 |
| hole effective mass : in units of m_0 , the free electron mass | 0.16, 0.49 | 0.044, 0.28 | 0.082, 0.45 |
| Intrinsic electron mobility at 300K ($\text{cm}^2\text{V}^{-1}\text{s}^{-1}$) | 1350 | 3900 | 8500 |
| Intrinsic hole mobility at 300K ($\text{cm}^2\text{V}^{-1}\text{s}^{-1}$) | 480 | 1900 | 400 |
| Electron diffusion coefficient at 300K (cm^2s^{-1}) | 35 | 100 | 220 |
| Hole diffusion coefficient at 300K (cm^2s^{-1}) | 12.5 | 50 | 10 |

2.3 Fermi Level in an intrinsic (undoped) semiconductor

If the material is undoped, then all the electrons in the conduction band (CB) must have been thermally excited from the valence band (VB). This fact is sufficient to tell us where the (E_f) should be. The electron and hole densities must be,

$$n = \int_{E_C}^{\infty} dE D(E) f(E) \quad (2.7)$$

$$p = \int_{-\infty}^{E_V} dE D(E) (1 - f(E)) \quad (2.8)$$

Let us assume that the dispersion relations are very simple

$$E_e(k) = E_C + \frac{\hbar^2 k^2}{2m_e} \quad (2.9)$$

$$E_h(k) = E_V - \frac{\hbar^2 k^2}{2m_h} \quad (2.10)$$

where E_C , E_V denote the bottom and the top of the conduction and valence bands respectively. The density of states (in 3D), including spin degeneracy is then given by:

$$D(E) = \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E - E_C)^{1/2} \quad \text{for} \quad E > E_C \quad (2.11)$$

$$D(E) = \frac{1}{2\pi^2} \left(\frac{2m_h}{\hbar^2} \right)^{3/2} (E_V - E)^{1/2} \quad \text{for} \quad E < E_V \quad (2.12)$$

PROBLEM : Calculate the density of states in 2D and 1D for parabolic bands. Explain why it is allright to take one of the limits to be infinity in equations 2.7 & 2.8, even though all bands have finite extents.

Now to evaluate equations 2.7 & 2.8 we proceed as :

$$\begin{aligned} n &= \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \int_{E_C}^{\infty} dE (E - E_C)^{1/2} \frac{1}{e^{\beta(E-E_F)} + 1} \\ &= \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \frac{1}{\beta^{3/2}} \int_0^{\infty} du \frac{u^{1/2}}{e^u e^{-\beta(E_F-E_C)} + 1} \quad \text{where} \quad u = \beta(E - E_C) \\ &= 2 \left(\frac{2\pi m_e k_B T}{h^2} \right)^{3/2} \left(\frac{2}{\sqrt{\pi}} \int_0^{\infty} du \frac{u^{1/2}}{e^u e^{\beta(E_C-E_F)} + 1} \right) \end{aligned} \quad (2.13)$$

$$(2.14)$$

Now we identify the integral within the brackets as a Fermi-Dirac integral, defined as :

$$F_j(z) = \frac{1}{\Gamma(j+1)} \int_0^{\infty} dx \frac{x^j}{e^z e^x + 1} \quad (2.15)$$

Further the "effective density of states" in the conduction band is defined as

$$N_C = 2 \left(\frac{2\pi m_e k_B T}{h^2} \right)^{3/2} \quad (2.16)$$

Note however that the dimension of N_C is not the same as $D(E)$. With the definitions eqn 2.15 & 2.16, eqn 2.13 then reduces to

$$n = N_C F_{1/2} \left(\frac{E_C - E_F}{k_B T} \right) \quad (2.17)$$

You can prove that the number of holes is given by

$$p = N_V F_{1/2} \left(\frac{E_F - E_V}{k_B T} \right) \quad (2.18)$$

The Fermi-Dirac integrals appear often in physics. They are tabulated as "special functions". We can show that if E_F is reasonably *below* E_C , such that $\frac{E_F - E_C}{k_B T} < -4$ the integral is very closely approximated by $e^{\frac{E_F - E_C}{k_B T}}$. This is called the non-degenerate regime where the electron and hole densities are given by

$$n = N_C e^{\beta(E_F - E_C)} \quad (2.19)$$

$$p = N_V e^{\beta(E_V - E_F)} \quad (2.20)$$

For charge neutrality we must have $n_i = p_i$ for undoped (intrinsic) semiconductors only. Multiplying eqns 2.19 & 2.20

$$\begin{aligned} n_i^2 &= n_i p_i = N_C N_V e^{\beta(E_C - E_V)} \\ \therefore n_i &= \sqrt{N_C N_V} e^{\beta E_g/2} \end{aligned} \quad (2.21)$$

Clearly the intrinsic carrier density falls rapidly with increasing band gap. We can compare the result with the data in table 2.1.

Note however that the product of the carrier densities is independent of the location of the Fermi level even when $n \neq p$. This is a very important fact and allows us to write

$$np = n_i^2 \quad (2.22)$$

even when the source of the charges are dopants. In such cases (we will see in the next section) the Fermi level moves away from its intrinsic position. The electron and hole densities can then become vastly unequal - but they do so in such a way that the np product still remains the same. We now solve eqns 2.19 & 2.20 for E_f and get the intrinsic Fermi level (E_{fi})

$$E_{fi} = \frac{E_C + E_V}{2} + \frac{3}{4} k_B T \ln \frac{m_h}{m_e} \quad (2.23)$$

PROBLEM : Show that the deviation of the electron density (n) from intrinsic density (n_i) and the deviation of the Fermi level (E_f) from the intrinsic Fermi level (E_{fi}) are related as :

$$n = n_i e^{\beta(E_f - E_{fi})} \quad (2.24)$$

2.4 Fermi level in a doped semiconductor

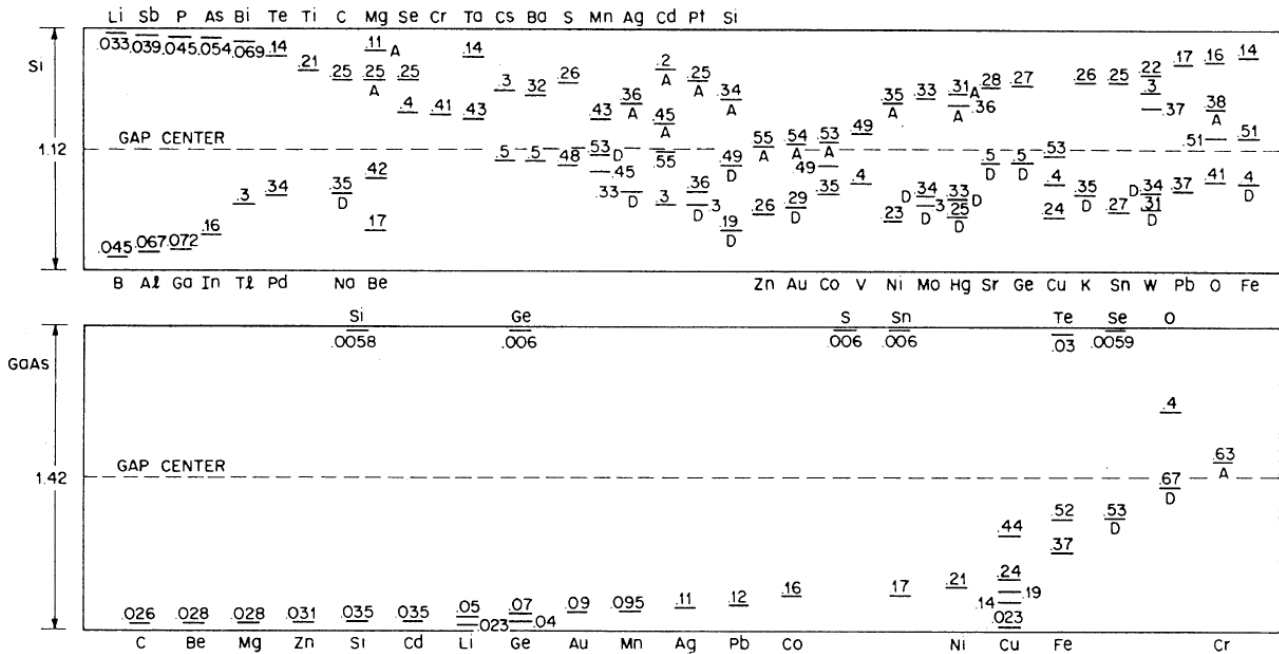


Figure 2.1: Impurity levels in Silicon and Gallium Arsenide (taken from the book by S.M. Sze). Notice that the shallowest levels tend to cluster somewhat around a value.

We now come to the more practical situation, where there are dopants and ask: where is the Fermi level? If there are dopants then n and p are no longer equal. In fact the number of carriers supplied by ionised dopants can be several orders larger than the intrinsic carrier densities.

The fundamental point is that all the atoms of the host lattice and the dopants were initially neutral. But inside the semiconductor there are now four sources of charge :

- Negatively charged electrons in the conduction band (n)
- Unoccupied (positively charged) ionised donor atoms (N_d^+)
- Negatively charged ionised acceptor atoms (N_a^-)
- Unoccupied states (holes) in the valence band

The sum total of all these four must continue to be zero. To start with the valence band was full and the conduction band was empty (intrinsic semiconductor), then we put in neutral donor atoms (capable of giving out an electron) and neutral acceptor atoms (capable of capturing an electron). So the sum total must remain zero.

Thus if we can write down the carrier concentrations in the conduction and valence band and calculate the fraction of dopants which are ionised (as a function of E_f) then we can have an equation where E_f is the only unknown. This is how one determines the location of E_f

$$n + N_a^- = p + N_d^+ \quad (2.25)$$

We know how to write n and p as a function of E_f using equations 2.19 and 2.20. Now we ask, what is the probability that a donor will ionise. This question is an interesting exercise in statistical physics. The donor site (*e.g.* Phosphorous in Silicon) can exist in four states

1. It may lose its electron (charge = +1 , energy = 0)
2. It may be occupied by a spin up electron (charge = 0, energy = E_D)
3. It may be occupied by a spin down electron (charge = 0, energy = E_D)
4. It may be occupied by one spin up and one spin down electron (charge = -1 , energy = $2E_D + U$ where " U " is the large repulsive energy cost of putting two electrons on the same site, making the state very improbable.)

The dopant densities are not very large compared to the density of atoms of the host lattice. It is rarely more than 1 in 10^3 to 10^4 . So we can treat each dopant atom in isolation and the electron can be localised on the atom.¹ Each dopant can exchange electrons with the "sea" of conduction band electrons. It is thus in equilibrium with a larger system and can exchange particles with it - thus its temperature and chemical potential must be the same as that of the larger system.

So we write the partition function as (with μ , the chemical potential set as E_f)

$$\begin{aligned} Z_G &= \sum_{E,N} e^{-\beta(E-\mu N)} \\ &= e^{-\beta(0-0)} + e^{-\beta(E_D-E_f)} + e^{-\beta(E_D-E_f)} + e^{-\beta(2E_D+U-2E_f)} \\ &\approx 1 + 2e^{-\beta(E_D-E_f)} \end{aligned} \quad (2.26)$$

The mean occupancy (probability that the dopant is *not* ionised) is then,

$$\begin{aligned} 1 - \frac{N_D^+}{N_D} &= 0.P(0) + 1.P(\uparrow) + 1.P(\downarrow) + 2.P(\uparrow\downarrow) \\ &= \frac{2e^{-\beta(E_D-E_f)}}{Z_G} \\ &= \frac{2e^{-\beta(E_D-E_f)}}{1 + 2e^{-\beta(E_D-E_f)}} \\ &= \frac{1}{\frac{1}{2}e^{\beta(E_D-E_f)} + 1} \end{aligned} \quad (2.27)$$

¹However if the dopant densities *are* very high then the dopant states will not be localised. This condition is called "Mott transition". It is among the most studied problems in semiconductor physics.

Note that this is not simply a Fermi-Dirac distribution. The fraction of ionised donors is

$$\frac{N_D^+}{N_D} = \frac{1}{1 + 2e^{-\beta(E_D - E_F)}} \quad (2.28)$$

The similar expression for the fraction of ionised (negatively charged) acceptors is

$$\frac{N_A^-}{N_A} = \frac{1}{1 + 4e^{-\beta(E_F - E_A)}} \quad (2.29)$$

The factor 4 is a result of the fact that the electron sitting on the acceptor could have come from four possible places - spin up/down from heavy hole band, spin up/down from light hole band. The split off band does not come into the picture because it is too far down.

One type of dopant only

If we neglect the valence band and the acceptors (which can be justified if only donors are present), combining eqns 2.19 and 2.28 we get

$$N_C e^{\beta(E_F - E_C)} = \frac{N_D}{1 + 2e^{-\beta(E_D - E_F)}} \quad (2.30)$$

E_F is the only unknown in eqn 2.30 and can be solved (numerically if required).

PROBLEM : Show that the carrier density can now be obtained by solving the following equation: (which is in turn obtained by using eqn 2.30)

$$n^2 + nN_C \frac{e^{-\beta\Delta}}{2} - N_D N_C \frac{e^{-\beta\Delta}}{2} = 0 \quad (2.31)$$

where $\Delta = E_C - E_D$.

The fermi level can be obtained by solving

$$x^2 + x \frac{e^{-\beta\Delta}}{2} - \frac{N_D}{N_C} \frac{e^{-\beta\Delta}}{2} = 0 \quad (2.32)$$

where $x = e^{\beta(E_F - E_C)}$

If you put $N_D = 0$ in either of the two equations you would get an unphysical answer. Why is this so?

PROBLEM :

In a system with N_D donors, N_A acceptors, N_D^+ donors and N_A^- acceptors are ionised. Each donor (acceptor) level has a degeneracy of g_D (g_A). There are n electrons in the conduction band and p holes in the valence band. (In general $g_D = 2$, but g_A may be different from 2.). Then

$$N_D^+ = \frac{N_D}{(g_D n / N_C) \exp \beta(E_C - E_D) + 1} \quad (2.33)$$

And the corresponding result for the acceptors:

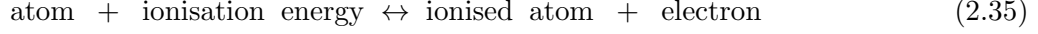
$$N_A^- = \frac{N_A}{(g_A p / N_V) \exp \beta(E_A - E_V) + 1} \quad (2.34)$$

Here N_C and N_V are the conduction and valence band effective density of states which have been defined earlier. Notice that the fermi energy does not appear in these relations.

A semiconductor may be doped with both (acceptors and donors) types of dopants. In a situation where there are a large number of donors and a few acceptors (*i.e.* $N_D \gg N_A$), how would eqn 2.31 (the previous problem) be modified?

2.4.1 Thermal ionisation (Saha equation) of the dopant system

It is instructive to calculate the fraction of ionised dopants in another way. We can think of the problem as a thermal ionization of bound states - in a way that is very similar to the method of calculating the ratio of ionised to unionised atoms (of a certain species) in a hot plasma. We want to find the "chemical equilibrium point" of the reaction:



A certain fraction of atoms will exist in the dissociated state and a certain fraction will remain in the undissociated state. The fraction which minimises the free energy of the entire system (at a certain temperature) will be the equilibrium point.

Taking this approach we can calculate the ratio N_D^+/N_D by minimising the free energy of the entire system of free electrons and the dopants. First we write the free energy so that the free electron concentration n is the only variable.

$$\begin{aligned} F_{\text{system}} &= F_{\text{electrons}} + F_{\text{dopants}} \\ F_{\text{electrons}} &= -kT \ln \frac{z^n}{n!} \text{ where for a single electron} \\ z &= \sum_{\text{all states}} e^{-\beta E} \\ &= V \frac{2}{h^3} \int d^3 \mathbf{p} e^{-\frac{\beta p^2}{2m}} \\ &= 2V \left(\frac{2\pi m kT}{h^2} \right)^{3/2} \end{aligned} \quad (2.36)$$

Now since $N_D - n$ dopant sites are occupied we have for the internal energy (U) and entropy (S)

$$U = -\Delta(N_D - n) \quad (2.37)$$

$$S = k \ln \left(2^{N_D - n} \frac{N_D!}{n!(N_D - n)!} \right) \quad (2.38)$$

$$F_{\text{dopants}} = U - TS \quad (2.39)$$

PROBLEM : Minimise $F_{\text{system}} = F_{\text{electrons}} + F_{\text{dopants}}$ w.r.t n , using Stirling's approximation for factorials as needed and show that you get exactly the same result as eqn 2.31. This is essentially a variant of the "Saha ionisation" equation, applied to a situation where the atoms and ions are not mobile, but only the electrons are.

2.4.2 General method of solving for the Fermi level

Consider a situation where a semiconductor is doped with N_D donors and N_A acceptors. We want the general solution for the location of E_F and all the carrier densities, ionisation probabilities.

Since the semiconductor is overall neutral we have using the charge neutrality condition

$$n + N_A^- = p + N_D^+ \quad (2.40)$$

$$N_C F_{1/2} \left(\frac{E_C - E_F}{k_B T} \right) + \frac{N_A}{1 + g_A e^{\beta(E_A - E_F)}} = N_V F_{1/2} \left(\frac{E_F - E_V}{k_B T} \right) + \frac{N_D}{1 + g_D e^{\beta(E_F - E_D)}} \quad (2.41)$$

- here $g_A = 4$ and $g_D = 2$ are the acceptor and donor degeneracies. E_A and E_D are the acceptor and donor levels.

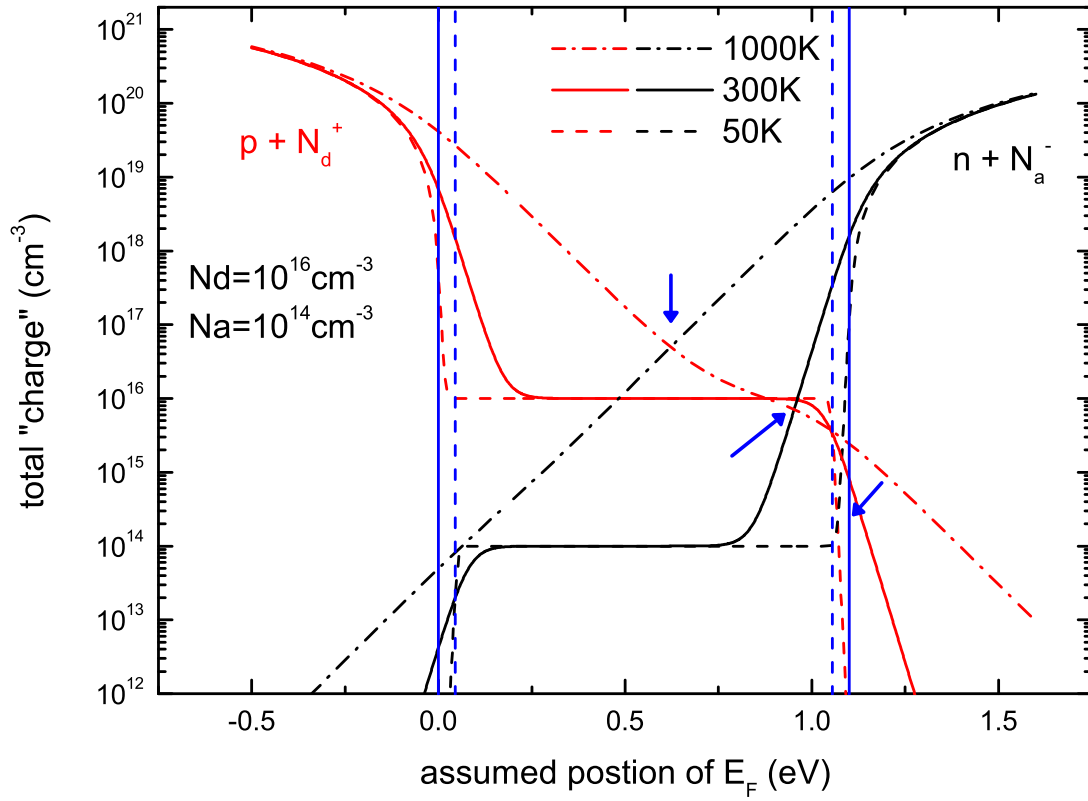


Figure 2.2: The LHS and RHS of eqn 2.41 are plotted for different temperatures. The intersection point is the solution for E_F .

- Since E_F is the only unknown here, we can plot the LHS and RHS by treating E_F as an independent variable. The position where they intersect must be the solution. Figure ... illustrates the situation.
- Notice that E_F is temperature dependent.
- Once E_F is determined all the quantities can be determined. In general this cannot be done analytically. An example on numerical solution is given in Fig 2.2, 2.3.

Figure 2.3: Figure shows the resulting carrier densities obtained from $n(T)$. Notice the various regimes.

2.5 Metal-semiconductor and semiconductor-semiconductor junctions

We now know the following important things :

1. How to calculate the charge density, if we know the location of E_f . Ignoring the holes and acceptors for the time being to keep the number of terms to a minimum, we have

$$n(x) - N_D^+(x) = N_C e^{\beta(E_f(x) - E_C(x))} - N_D \frac{1}{1 + 2e^{-\beta(E_D(x) - E_f(x))}} \quad (2.42)$$

$$\rho(x) = -|e|(n(x) - N_D^+(x)) \quad (2.43)$$

2. The charge density is related to the electrostatic potential (V) as

$$\nabla^2 V(x) = -\frac{\rho(x)}{\epsilon_r \epsilon_0} \quad (2.44)$$

3. The scalar potential is essentially the bottom of the conduction band.
4. In equilibrium E_f is constant, recall that current flow requires a gradient in the electrochemical potential or Fermi level.

2.5.1 Situations with no current flow

Now let's see how we can put this in practice - a (somewhat idealised) metal in contact with a semiconductor (see fig). The work function of a metal (ϕ_m in our discussion) is the energy an electron sitting at the Fermi level of the metal needs to escape from inside the metal to outside (vacuum level).² Similarly ϕ_s is the work function of the semiconductor in question. The two objects are brought in contact, so that they can exchange electrons. If $|\phi_s| < |\phi_m|$, then transferring an electron from the semiconductor to the metal is energetically favourable.

There is a little complication though - in a semiconductor the Fermi level is often in a gap - thus no electron may actually be right at the Fermi level. To account for this we define the electron affinity (χ) of the semiconductor as the energy difference between the vacuum level and the bottom of the conduction band of the semiconductor sufficiently deep inside.

When the two objects touch the conduction band and the Fermi level of the metal would be separated by $\phi_B = \phi_m - \chi$. Since deep inside the semiconductor (where the surface should have no effect) the bottom of the band and the Fermi level must continue to be separated by $\phi_s - \chi$, this dictates that the total drop of the conduction band of the semiconductor must be $\phi_m - \phi_s$.

Charge separation must give rise to an extra electrostatic potential - and it is reasonable to expect that the bands would start bending in a way that would result in a barrier preventing the flow after some time. At this point the metal and the semiconductor's Fermi level must be identical. Applying the set of conditions that we just talked about produces the band diagram shown in Fig. 2.4.

Some charge has moved from the semiconductor to the metal. This charge came from the dopants sitting considerably above E_f . If a dopant site is pushed above E_f then it must be charged, because the electron cannot reside at a site sitting much above E_f . The bands in the metal didn't have to bend a lot to accomodate this extra charge, because the density of states of a metal near E_f is very large.

To (numerically) solve eqns 2.42, 2.43 and 2.44 we can proceed as follows.

²This can be typically about 4-5 eV, it depends a lot on which crystal face we are considering and how clean the surface is. Since we are going to ignore these aspects to highlight the basic concept, our discussion is a bit idealised here.

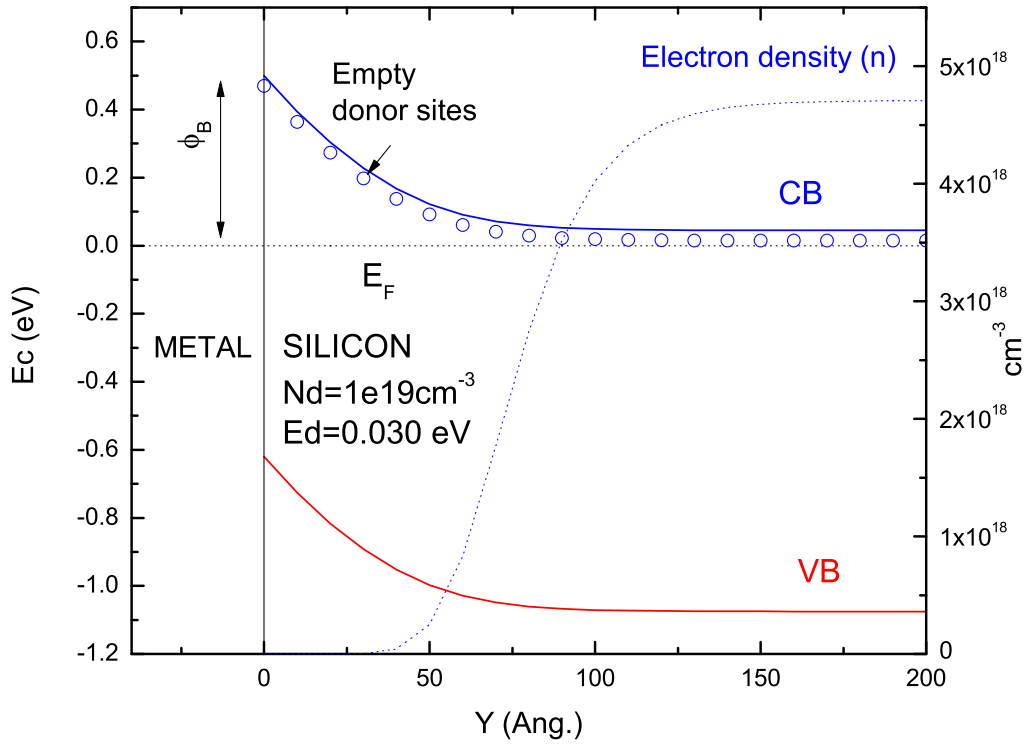


Figure 2.4: The band bending near the surface of a n-type semiconductor to metal contact, where $|\phi_s| < |\phi_m|$. The calculation has been done using a program written by Greg Snider (Notre Dam University)

1. Since E_f is constant, everything can be measured relative to E_f , by setting $E_f = 0$.
2. The gradient of the scalar potential is the same as the gradient of the conduction band bottom (E_C).
3. We make a guess for $E_C(x)$ and use this to calculate the expected charge density by using eqns 2.42 and 2.43.
4. This calculated charge density should give a new guess for the potential via Poisson's equation (eqn 2.44).
5. We use this potential and go back to step 3.
6. The iterative process can continue till the change in two successive iterations becomes very small (our convergence criteria)
7. But all these equations are differential equations - they need proper boundary conditions. In the calculation of Fig. 2.4, we set the slope $dE_C/dx = 0$ deep inside the material and $E_C = \phi_B$ at the other end. Choosing the correct boundary condition depends on the physical situation.

PROBLEM : Try to draw the band diagram of the metal-semiconductor contact when $|\phi_s| > |\phi_m|$ and the semiconductor is p-type. Where will the semiconductor accommodate the electrons flowing in?

Why doesn't the depletion zone extend to the metal as well?

We now apply the same process to a p-n junction, see Fig. 2.5.

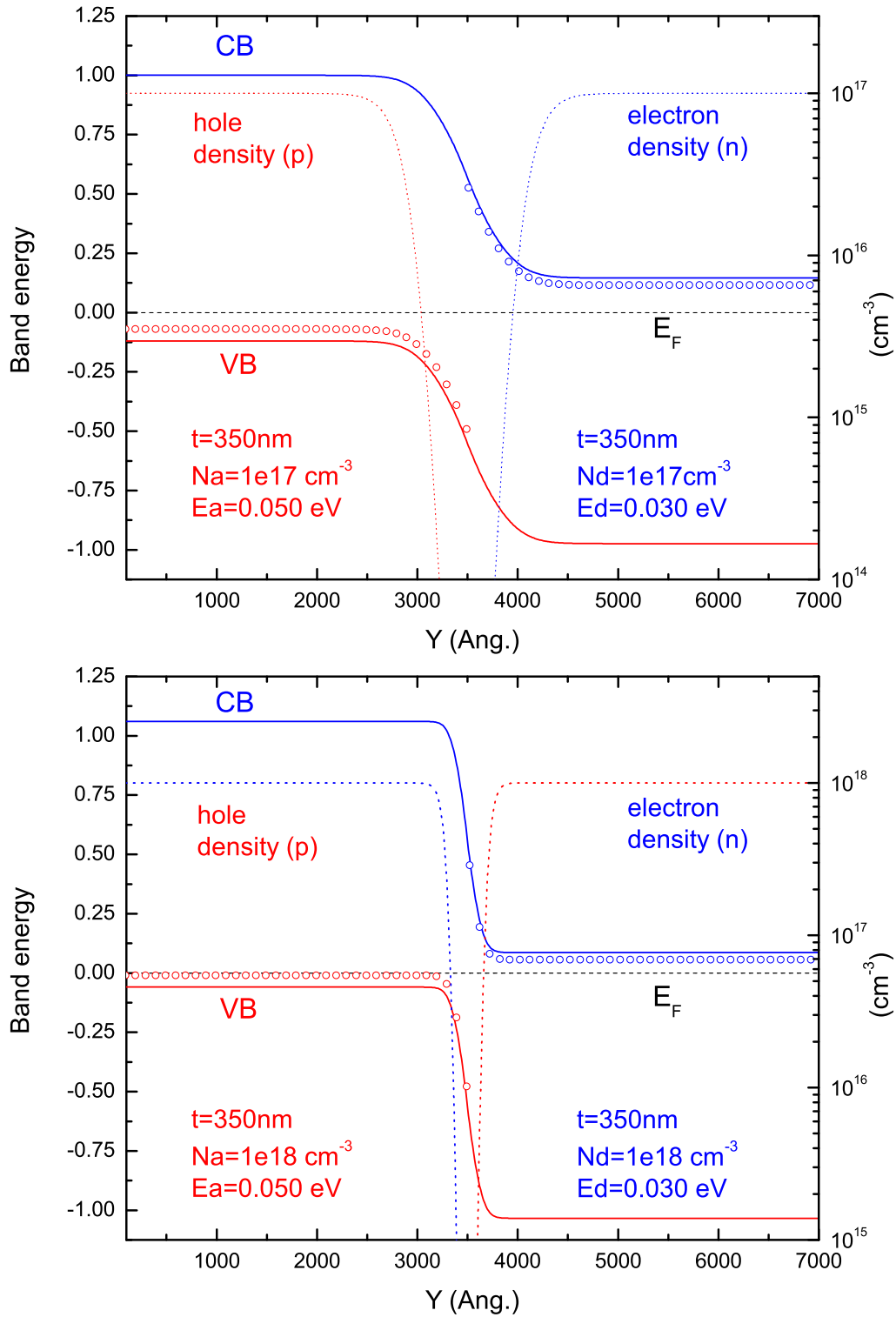


Figure 2.5: The band bending near a p-n junction. Note that the junction becomes sharper at a higher doping level. The Fermi energy also moves closer to the dopant level at higher doping concentration. The calculation has been done using a program written by Greg Snider (Notre Dam University)

2.5.2 How realistic are these calculations?

We remarked at the beginning of this section that there are some idealisations. The work function of a metal in reality depends on which crystal face we are using, how clean it is etc. This means that if we deposit a thin film of a metal (say gold on silicon) on a semiconductor, we can't really take the values for a crystal of gold and clean silicon and predict what the barrier will be. Also the density of states near the surface of a semiconductor is modified by the presence of surface states - which ultimately mean that the Schottky barrier needs to be determined experimentally. However the band diagram of the barrier that we drew and the principles for solving the band-bending are sufficiently generic.

2.5.3 When is a contact not a "Schottky" ?

In the previous section we considered the work function of the metal to be larger $|\phi_s| < |\phi_m|$ and the semiconductor to be n-doped. As a consequence some electrons flowed from the semiconductor to the metal. What if $|\phi_s| > |\phi_m|$. Clearly the band bending must be different, because if electrons flow into the semiconductor then the dopant states and the conduction band cannot remain much above the Fermi-level. But the dopants (if they drop below E_f must hold on to their own electrons (they must be occupied). and the conduction band would have to accommodate the electrons. Under these conditions no depletion zone can form and hence there should be no Schottky barrier.

Real ohmic contacts

In reality ohmic contacts on a semiconductor are made by depositing an alloy that often contains one noble metal (Gold) and another element that can act as a dopant. For example an alloy of Gold-Germanium is commonly used to make ohmic contacts to n-type Gallium Arsenide. After depositing the metal the sample is generally annealed (heated to a high temperature) very rapidly so that the Germanium diffuses into the surface, heavily dopes the region around it allowing the Gold to make a contact with no barrier. The microscopic mechanisms of ohmic contact formation are not very simple and you would find a good deal of research work happening on these.

Gold-Beryllium alloy can be used to contact p-type Gallium Arsenide.

Gold-Antimony alloy can be used to contact n-type Silicon...and so on.

2.5.4 Situations with varying E_f : what more is needed?

We noted earlier that the current is related to the gradient of the electrochemical potential (in a 1-dimensional case) as

$$j = -n(x)\mu \frac{dE_f(x)}{dx} \quad (2.45)$$

So we now have three variables to deal with - other than the charge density and the profile of the bottom of the band. We also need to calculate the the profile of $E_f(x)$.

We expect μ to be a function of n .

In general the product $n\mu$ would increase with increasing carrier density, it doesn't necessarily imply that μ will be larger at higher densities. However at least in a 1-dimensional situation it is easy to see that wherever $n\mu$ is large, $\frac{d\mu}{dx}$ must be small. This reminds us of what to expect if we apply a voltage across a string of resistances (in series). The largest voltage drop must occur across the largest resistance, because the current through each of them is constant.

When the current flow is very small we can approximate the situation by saying that all the drop in E_f must be across the most resistive region (like a barrier) if we can identify one. This is however an approximation to get around the fact that the variation of μ with n is in general a hard and very system dependent problem. An empirical approach is shown in Fig 2.6.

Mobility model

Empirical relations and estimates can be used to get mobility as a function of carrier density from experimental data, here's an example

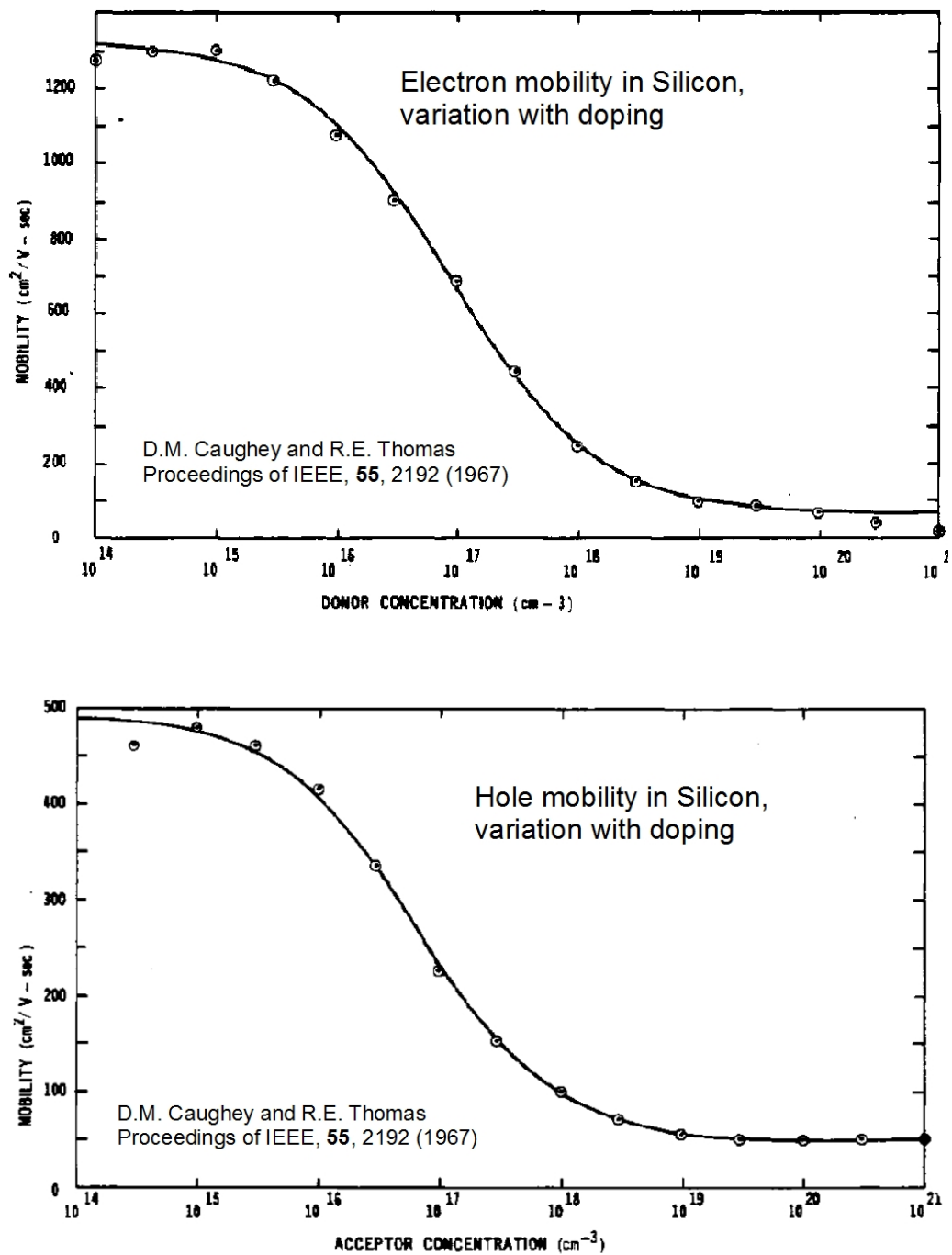


Figure 2.6: The variation of mobility with doping and carrier density can be empirically modelled from experimental data and used to solve the current equation numerically.

PROBLEM : Band bending at the p-n junction. The total drop in the profile of the bands shown in Fig. 2.5 can be calculated in two different ways. First let us see the method given in most text books. The flow of charge through the junction can be thought to have a drift (forced by the electric field) and a diffusion (forced by density gradient) component - at equilibrium, when the electrochemical potential is constant, these two components must add up to zero. So we get for no electron current

$$\begin{aligned} J_{drift} + J_{diffusion} &= 0 \\ -ne\mu \frac{dV}{dx} - De \frac{dn}{dx} &= 0 \end{aligned} \quad (2.46)$$

We have used the standard relation between current, diffusion constant and density gradient. (A full justification of this set of equations requires the Boltzmann transport formulation, which we haven't done.) Then solve the differential equation using $D/\mu = kT/e$ and the assumption that all dopants are ionised. So that the electron density on the n-side is $n = N_D$ and on the p-side it is $n = n_i^2/N_A$. You should get the result for the total change in electrostatic potential as one moves from one side of the junction to the other. The electron bands are higher on the p-side.

$$\Delta V = \frac{kT}{e} \ln \frac{N_A N_D}{n_i^2} \quad (2.47)$$

Now think of the same in another way. Let us not mention diffusion constant at all, but use the fact that the electrochemical potential (E_f) is constant. Here the free energy of the electrons can be written, including the electrostatic potential as

$$F = -k_B T \ln \frac{z^n}{n!} + neV \quad (2.48)$$

where the electron density $n(x)$ is a function of position. And

$$z = 2\Omega \left(\frac{2\pi m k_B T}{h^2} \right)^{3/2} \quad (2.49)$$

is the partition function of a single free electron moving in the conduction band. Ω is the volume which should drop out of the calculation. Since we assume full ionisation we can neglect the entropy contribution coming from possible number of ways to distribute the bound electrons among the dopants.

Differentiating this w.r.t. n to get the electrochemical potential, first write

$$E_f(x) = \frac{\partial F}{\partial n(x)} \quad (2.50)$$

And then show that setting $E_f(x) = \text{constant}$ leads to exactly the same condition as before. Convince yourself that in both cases the approximations that we made are actually identical. They are both consequences of Boltzmann statistics applied to the free electron gas in the conduction band.

The reverse and forward biased metal-semiconductor junction

In Fig. 2.4 we plotted the band diagram of a metal-semiconductor junction with no voltage applied ($E_f = \text{constant}$). No current flows at equilibrium. The tunnel rates from both sides balance each other. Now imagine that the electron energies on the metal side is raised by connecting the metal to the negative terminal of a battery. The drop in the electrochemical potential must happen predominantly over the depletion region. See Fig. 2.7

- Electrons which try to cross over from the metal to the semiconductor still see almost the same barrier. The current that can pass through is $I_{m \rightarrow s} = AT^2 e^{-\phi_B/kT}$, where A is a constant.

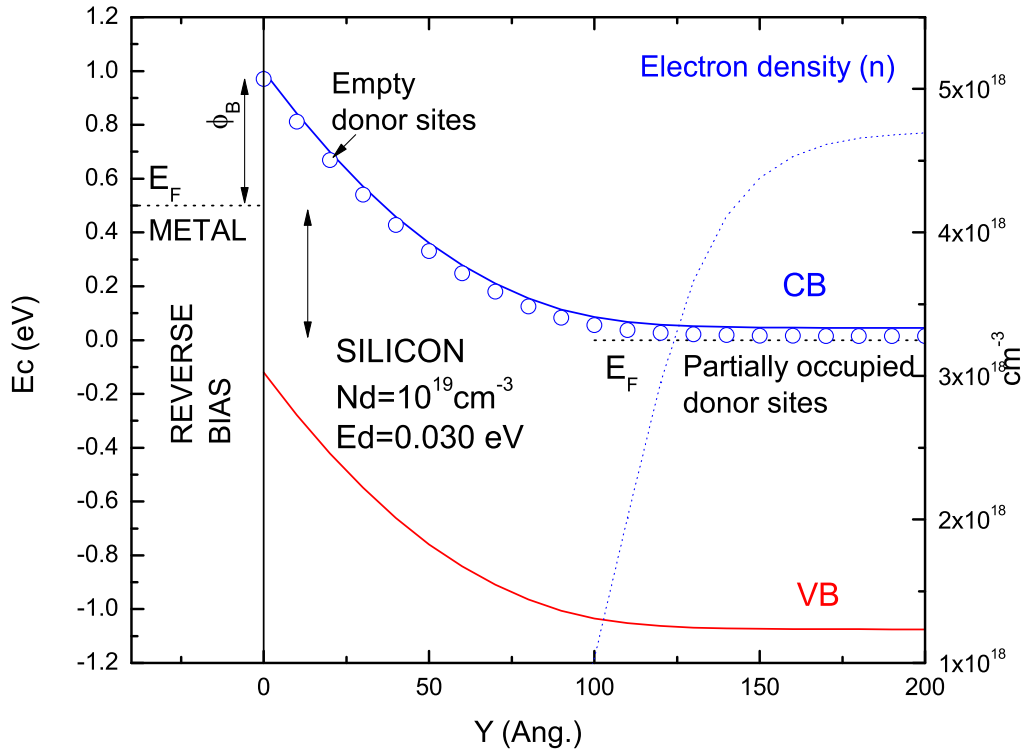


Figure 2.7: Approximate band bending near a reverse biased metal-semiconductor junction. The calculation has been done using a program written by Greg Snider (Notre Dam University)

- But those which try to cross from the semiconductor to the metal now see a higher barrier. Typically tunnelling probability through a barrier would drop exponentially with the height of the barrier. $I_{s \rightarrow m} = AT^2 e^{-(\phi_B + V)/kT}$, where V is the voltage bias on the semiconductor w.r.t. the metal. In this case $V > 0$. Remember that positive voltage bias lowers electron energies.
- Thus only the reverse saturation current now flows

What happens when the electron energies in the metal are lowered? See Fig. 2.8

- Electrons which try to cross over from the metal to the semiconductor still see almost the same barrier. $I_{m \rightarrow s}$ remains the same.
- But those which try to cross from the semiconductor to the metal now see a lower barrier. Typically tunnelling probability through a barrier would increase exponentially as the height of the barrier is lowered. $I_{s \rightarrow m} = AT^2 e^{-(\phi_B - V)/kT}$
- A large number of electrons can now flow from the semiconductor to the metal. We take the difference of the left going and the right going currents to get the total current which is the well known diode equation : $I = I_0(e^{eV/kT} - 1)$, where I_0 , the reverse saturation current, is determined by the height of the Schottky barrier.

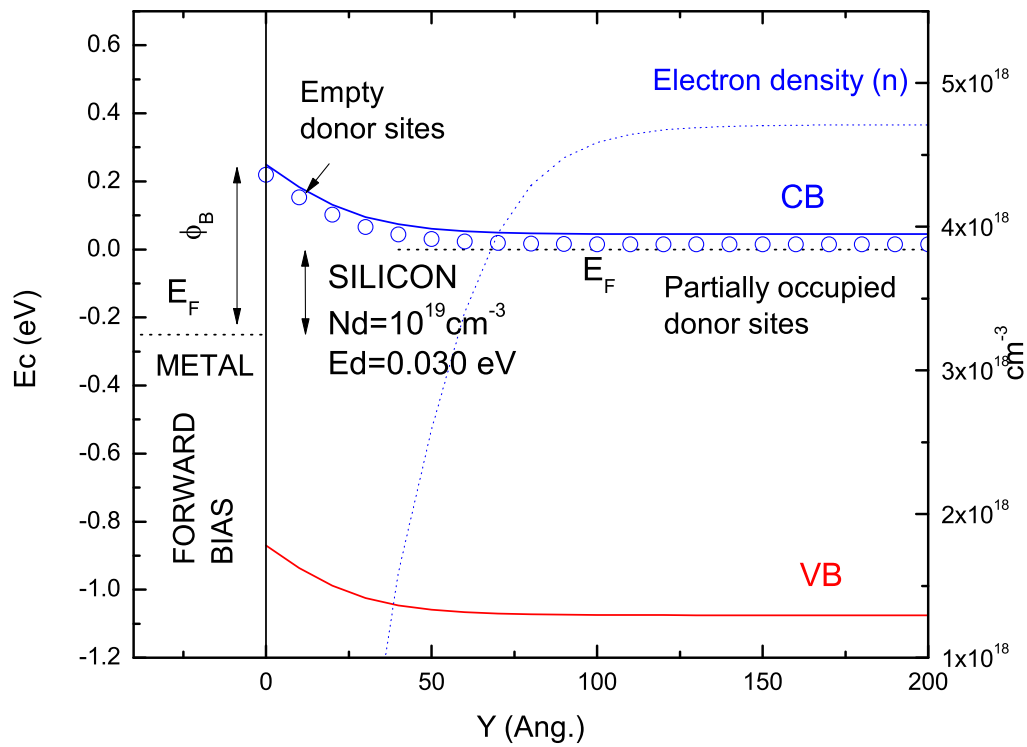


Figure 2.8: Approximate band bending near a forward biased metal-semiconductor junction. The calculation has been done using a program written by Greg Snider (Notre Dam University)

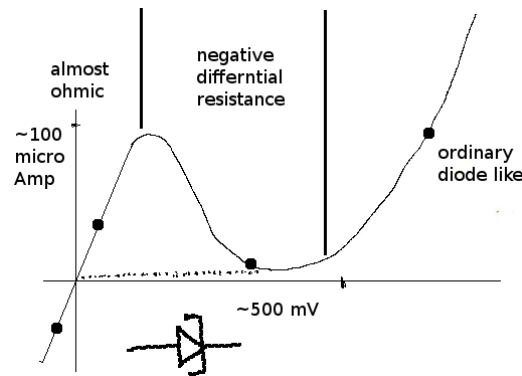


Figure 2.9: The difference in behaviour is at the low bias region, below what would be a typical cut-in voltage of the diode - typically $\sim 0.6-0.7$ V

2.6 The tunnel diode

We saw in the previous sections that the depletion zone separating the p-n sides gets narrower as the doping levels are increased. This is because the larger doping concentration allows the storage of larger amount of charge in the same thickness and hence allows sharper band slopes.

But this is not the only effect of increasing the doping. There can be a point after which the electrochemical potential itself moves into the conduction band and the doped semiconductor behaves almost like a metal. Of course for this to happen, the discrete energy level of the bound state would spread out into a continuum (like a band) that overlaps with the conduction band. What happens if we form a diode using very highly doped p and n sides? The answer is non-trivial.

Of course the structure forms a diode, but its characteristics are quite different at low bias. See the schematic of what happens first in Fig. 2.9.

1. First, at zero (or very small) bias the structure behaves just like a resistance. Notice the band structure shown in Fig. 2.10, at zero (or very small) there can be transfer of carriers, by tunnelling. This is because there are plenty of filled levels in the n-side which are aligned with the empty levels (holes) on the p-side. So the necessary conditions for tunnelling (thin barriers and equal energy filled and empty states) are met.
2. Now let us increase the forward bias by say about 100 mV, keeping it sufficiently below the cut-in voltage of a typical junction. The drop in the electrochemical potential would take place over the depletion region, and remain almost flat beyond that. But now the filled states on the n-side are getting aligned with the unfilled states on the p-sides which are close to the top of the valence band. The density of states near the end of the band is small. This would start to prevent tunnelling.
3. If the bias is increased still more the bottom of the CB on n-side will move past the top of the VB on the p-side. At this point no filled states are lined up with any empty state. This means that there can be very little current flow.
4. This is a strikingly different behaviour, compared to the ordinary diode. In fact the current will fall as the voltage is increased - that means the *differential resistance is negative*. This is the so called NDR region - which makes the tunnel diode special.
5. Now let us increase the voltage even further, so that the barrier in the CB itself comes down. From this point onwards it is an ordinary diode, in fact all the interesting things in a tunnel diode happens between zero and ~ 500 mV.

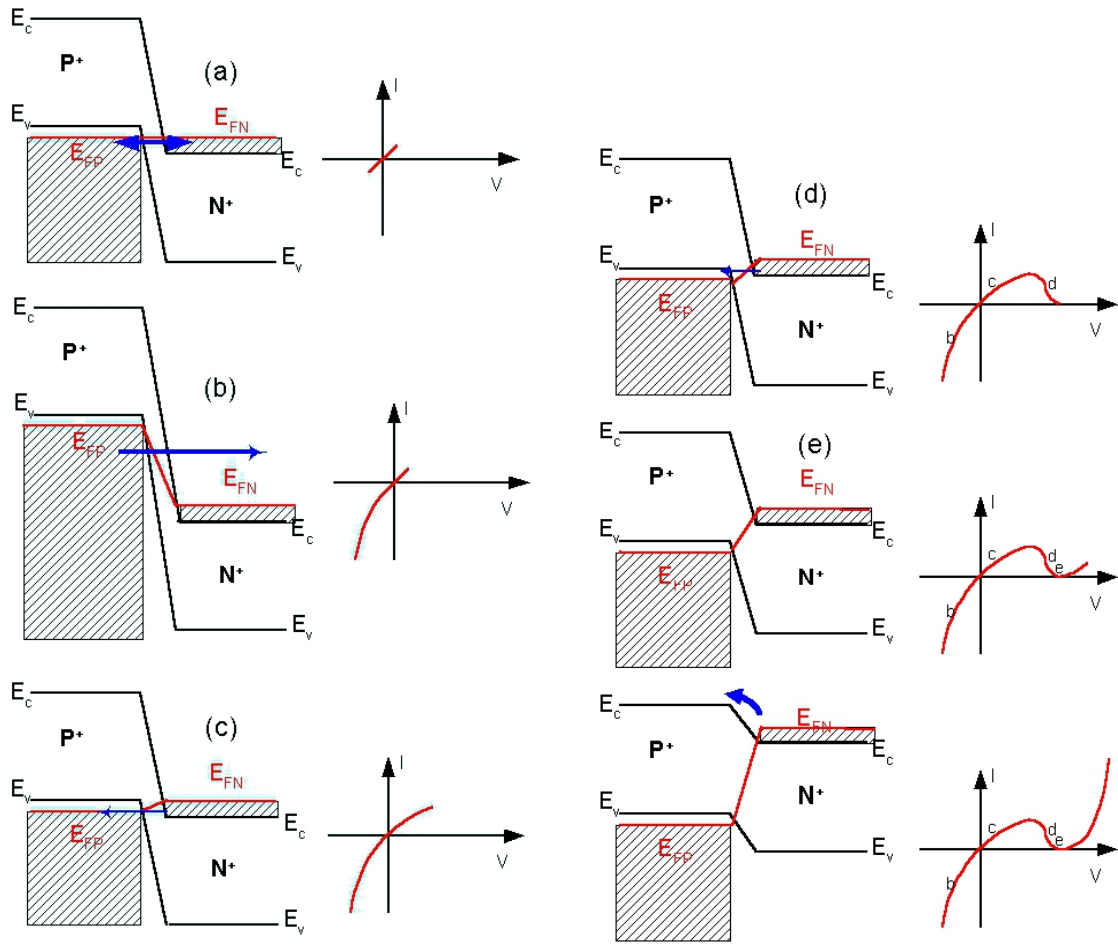


Figure 2.10: The band structure and the relative location of the Fermi energy at various points of the tunnel diode's I-V curve

6. The tunneling process, which distinguishes this device is also very *fast*. This is not obvious at first glance ! In fact the tunnel diodes can be used to make remarkably simple oscillators working at a few GHz. Historically this was the first solid state device to be able to do so.

Now consider the simple circuit block shown in the Fig 2.11. We will claim that it will oscillate if the dc bias point is chosen to lie in the NDR region. Let us follow the line of argument carefully

1. When it is first switched on the voltage starts increasing across tunnel diode + inductor. At this point because the TD is just like a resistance and the transient behaviour is that of an ordinary LR circuit.
2. Then the voltage tries to cross the turning point (peak in the I-V), and move down the curve. The current tries to decrease and hence the $L \frac{dI}{dt}$ drop across the inductor pulls the lower end of the TD to negative values. This means that the total drop across TD is now larger than the peak voltage.
3. But the I-V of the TD at large voltages must jump to the rightmost section (normal diode-like). That is the only way to satisfy Kirchoff's laws and the TD's characteristic.
4. But this is not a stable state. If the current tried to remain steady it cannot - because then the drop across the inductor is zero and the bias is not sufficient to hold it there. The current cannot continue to increase either - that would make the drop across TD+L larger than the bias voltage too!

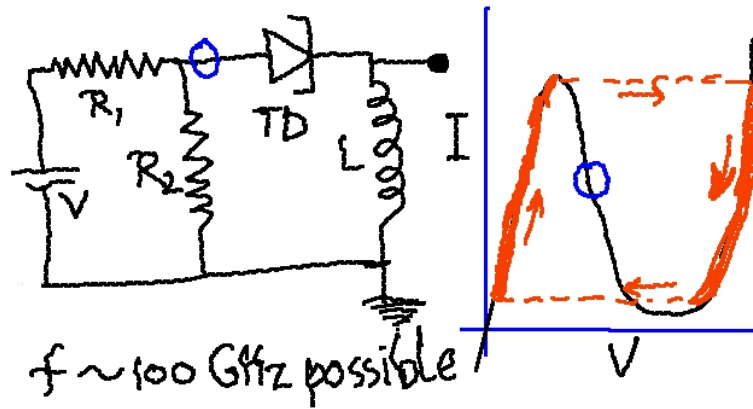


Figure 2.11: A simplified schematic of a tunnel diode oscillator. The bias point is in the NDR segment of the I-V curve. This is done by choosing V , R_1 and R_2 .

5. The TD must now move towards the bias point by sliding down along the curve. It does it till such time it again encounters a turning point, where it touches the bottom of the valley.
6. By a similar logic we can show that the only possibility is to jump back to the initial section of the I-V curve, near zero. The whole cycle will now repeat.

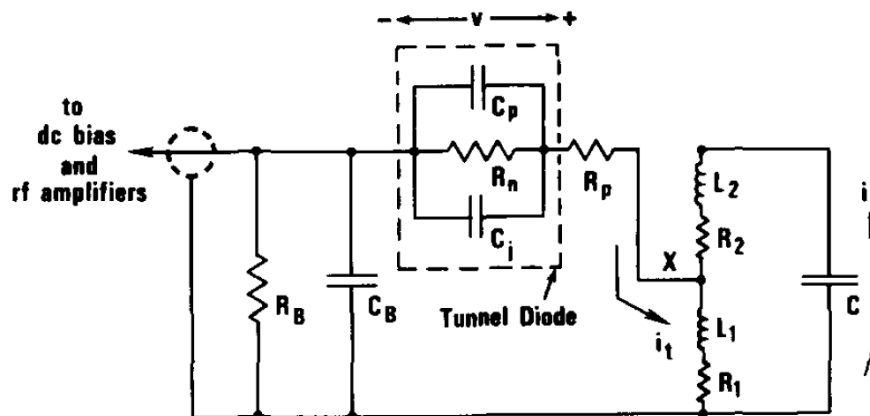


Figure 2.12: An equivalent circuit for tunnel diode. The figure is taken from: "Modeling of tunnel diode oscillators" Craig T. Van Degrift and David P. Love *Review of Scientific Instrumentation*, **52**, 712 (1981)

But why is this process very fast? Let's compare this circuit with another very simple oscillator made out of connecting the output of a NOT-gate to its own input! This also oscillates, but the fastest oscillation is determined the propagation delay through the transistors of the NOT-gate. This is a drift-diffusion driven process, not a tunneling process. The oscillations of the operating point of the TD can occur much faster, because it does not involve the full passage of a stream of electrons from the input to output - that is necessary for a thermionic valve or a transistor to work.

Fig. 2.11 is a bit oversimplified, a realistic model would have to include junction capacitances of the diode itself and the connecting coaxial cable etc. All realistic models of high frequency circuits need to include these. The oscillation frequency would be set by the LC combinations present in the circuit, see Fig. 2.12

2.6.1 The Bipolar Junction Transistor

We know how a BJT is biased and used, let's look at its band diagram and review a few salient points.

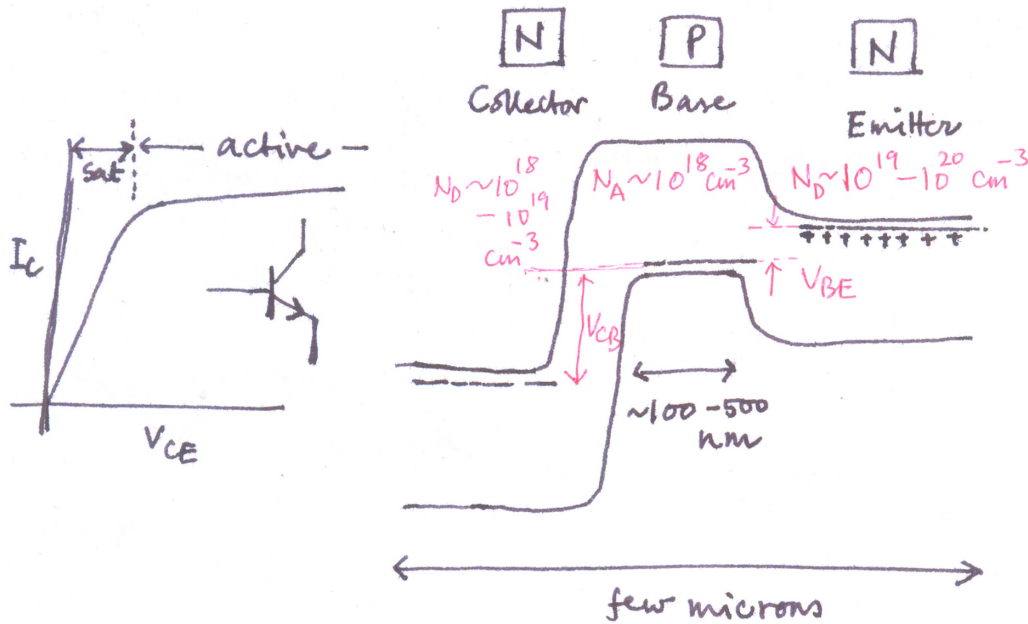


Figure 2.13: Band bending in a *npn* bipolar junction transistor under normal bias conditions. Notice the typical $I_C - V_{CE}$ curve of a transistor for a fixed I_B . How does the band profile shown lead to that kind of IV ?

- First notice the typical size and doping levels given in the schematic Fig 2.13. The base-emitter junction is forward biased, the base-collector junction is reverse-biased. Notice that the base is very thin - it is an essential feature. Why?
- See the following data (INSPEC database) for lifetimes and diffusion Lengths of *Minority Carriers* in Silicon

| N_D (cm^{-3}) | lifetime (sec) | Diffusion length (cm) | N_A (cm^{-3}) | lifetime (sec) | Diffusion length (cm) |
|-------------------------------|-------------------|-----------------------------|-------------------------------|-------------------|-----------------------------|
| 1.60E+21 | 2.164E-12 | 1.708E-06 | 1.20E+21 | 7.095E-12 | 2.679E-06 |
| 7.38E+19 | 9.635E-10 | 3.029E-05 | 1.70E+20 | 3.001E-10 | 2.565E-05 |
| 4.53E+18 | 1.384E-07 | 5.136E-04 | 8.49E+18 | 2.767E-08 | 3.144E-04 |
| 7.84E+16 | 1.606E-05 | 9.687E-03 | 2.77E+17 | 1.038E-06 | 4.630E-03 |

- The width of the base is comparable with the diffusion length - this is critical to its operation. Consider an electron injected through the emitter. The forward biased emitter junction injects electrons into the p-type base. A small fraction (often $< 1\%$) recombines and exits through the base lead. The fraction which passes through is denoted by $\alpha = \frac{I_C}{I_E} \approx 1$
- The rest shoots through and drops into the collector. Notice that once minority carriers are somehow injected into the base, it only needs to "fall" in energy as it moves towards the collector. Irrespective of how large the V_{CB} reverse bias is, the current will remain almost the same over

a large range - this is the active region of the transistor. The collector current is thus *almost entirely* fixed by the V_{BE}

$$I_C = \text{constant} \cdot \left(e^{\frac{eV_{BE}}{kT}} - 1 \right) \quad (2.51)$$

This equation is called the Ebers-Moll equation. I_C is thus very sensitive to V_{BE} and also to temperature. Since V_{BE} is a forward bias, it cannot be very different from a typical diode-drop, which is ≈ 0.7 volts for Silicon. However just like the massive gain of an op-amp is not generally used directly, this sensitive dependence is also not used directly. Usually feedback or some other aspect of design will have to ensure that sensitivity does not lead to instability.

- Because of the extra resistive drops in the body of the emitter, it is difficult to accurately determine the true base-emitter junction voltage. But I_B can be measured with certainty. So this is commonly used as the variable parameter rather than V_{BE} . Obviously $I_C = \frac{\alpha}{1-\alpha} I_B$. This ratio I_C/I_B is usually called β or h_{FE} and may have a typical value around 100. The subscript FE denotes "Forward biased Emitter".
- It should be clear to you at this point that two back-to-back diodes don't make a transistor. The flight of the injected carriers through the base and its drop into the collector would not happen if the base is "thick".
- However as V_{CB} increases, the depletion region penetrates more and more into the base. There will be a point when the entire base region will be depleted - this "punch-through" sets the maximum voltage the transistor can handle.
- Since the carriers have to diffuse through the base region, its time of passage sets the maximum frequency at which transistor action is still possible.
- The maximum dissipation happens at the CB interface. That is why the collector needs to be heat sunk properly. In metal-bodied packages and power transistors, the large metallic body is in good thermal contact with the collector. You should be able to understand why "body-collector notch-emitter" is typical for metallic transistor packages.
- Finally you would notice that the collector doping is less than the emitter doping. This ensures that the reverse biased CB junction does not see a very narrow depletion region where the field can become too strong.
- If you connect a transistor with its collector and emitter terminals interchanged, then it would still work as a transistor, but not a well designed one.

Chapter 3

MOSFETs and GaAs-AlGaAs heterostructures : accumulation of the 2-dimensional electron gas

References:

1. Chapter 8 Sec. 3 (4th edition), *Solid State Electronic Devices*, B. G. Streetman
 2. Chapter 9 *Physics of low-dimensional semiconductors* J.H. Davies
 3. Greg Snider's homepage has the tool used to calculate band structures.
See <www.nd.edu/~gsnider>
-

3.1 The MOSFET

The Metal-Oxide-Semiconductor Field effect transistor is like a capacitor. One plate is a Silicon substrate(p-doped for n-channel MOSFET - it will become clear why) itself. The dielectric is 10-100nm of SiO₂. The other plate is a layer of metal (the gate). Electrical contacts are made by implanting necessary dopants at selective places (see Fig/ any book on semiconductor devices).

In the band structure examples that follow, it was assumed that the acceptor concentration is $N_a = 1 \times 10^{17} \text{cm}^{-3}$. Notice how the bands progressively bend downwards, finally allowing enough electrons to accumulate in an originally p-type substrate. The Fermi level on two sides of the oxide are taken to be constant. This works because of the very high resistance of the oxide layer, the "drop" must happen across the oxide. The exact details of how it drops is not important because the oxide is devoid of electrons. For the purpose of the simulations, we will assume that the drop is linear.

This obviously means that the MOSFET's terminals are connected to a battery, whose voltage is being increased in steps, as we track the bending of the bands.

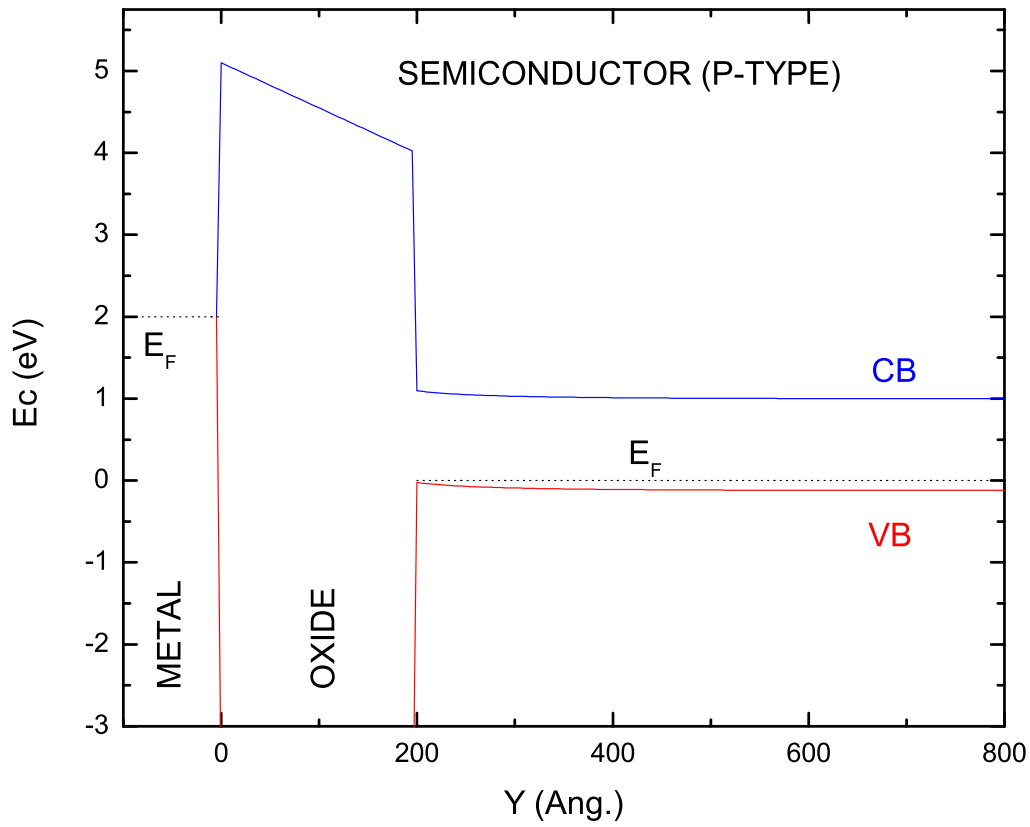


Figure 3.1: MOSFET in Accumulation region. Gate is negatively biased w.r.t the silicon and attracts more holes

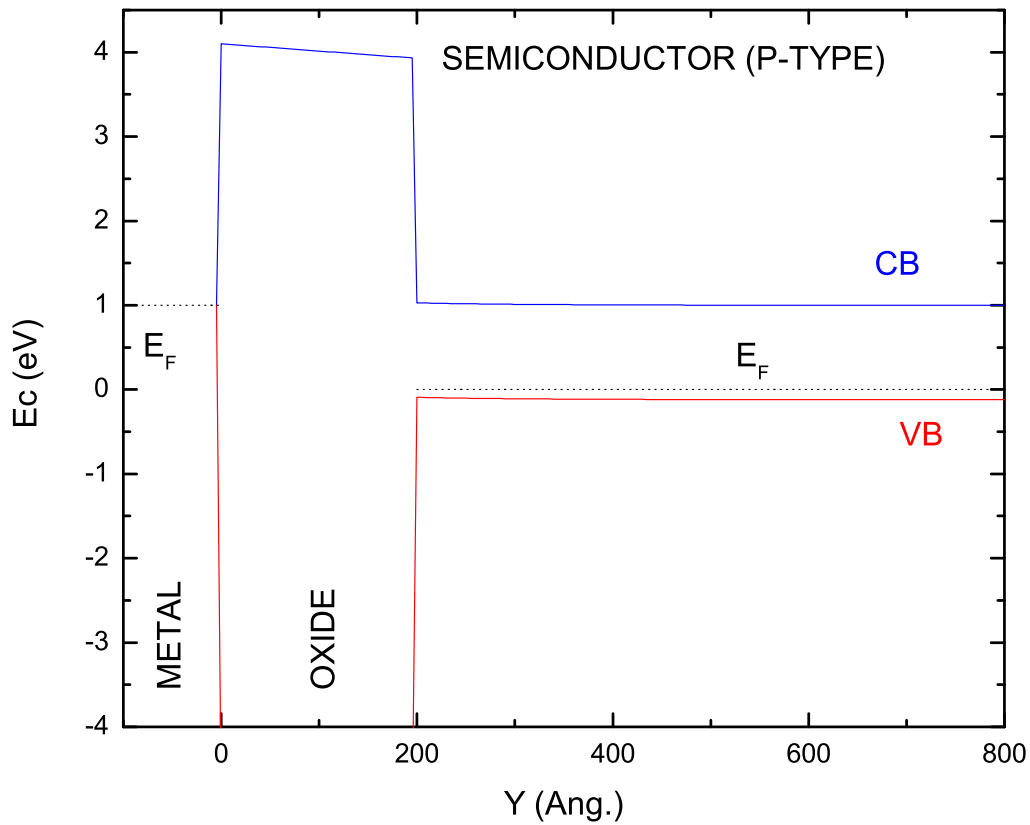


Figure 3.2: MOSFET in Accumulation region. Gate is negatively biased w.r.t the silicon and attracts more holes. the parameters have been chosen to give a near flatband condition

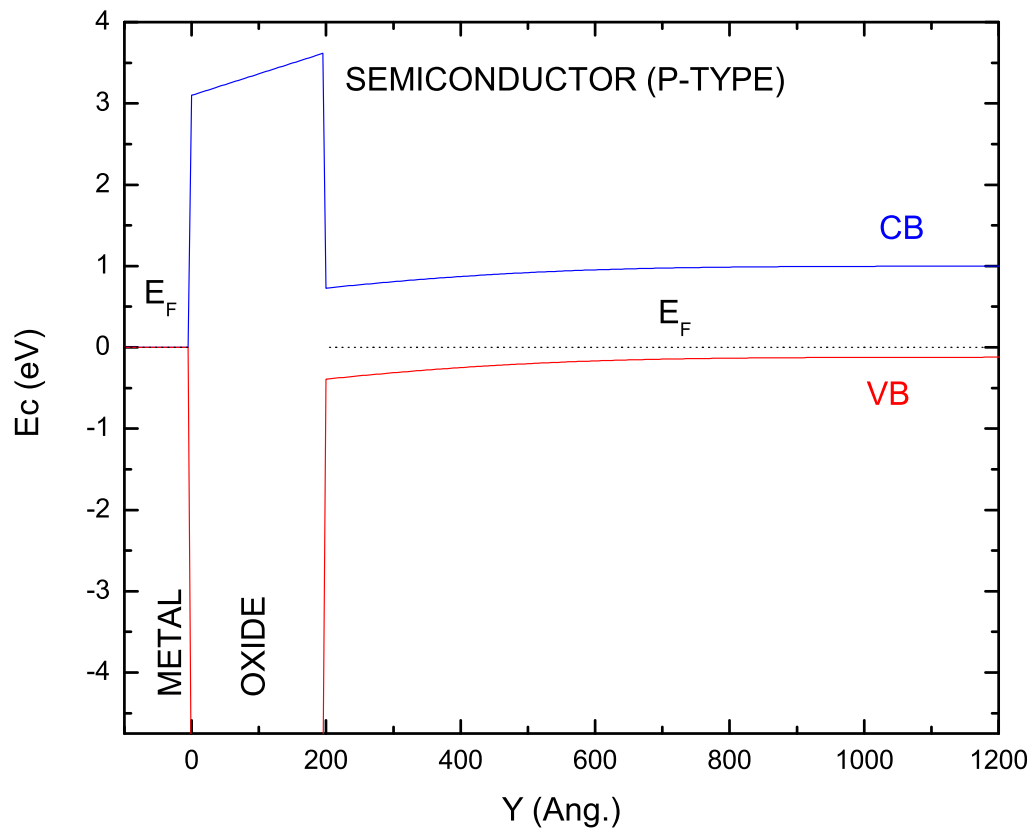


Figure 3.3: MOSFET with no bias.

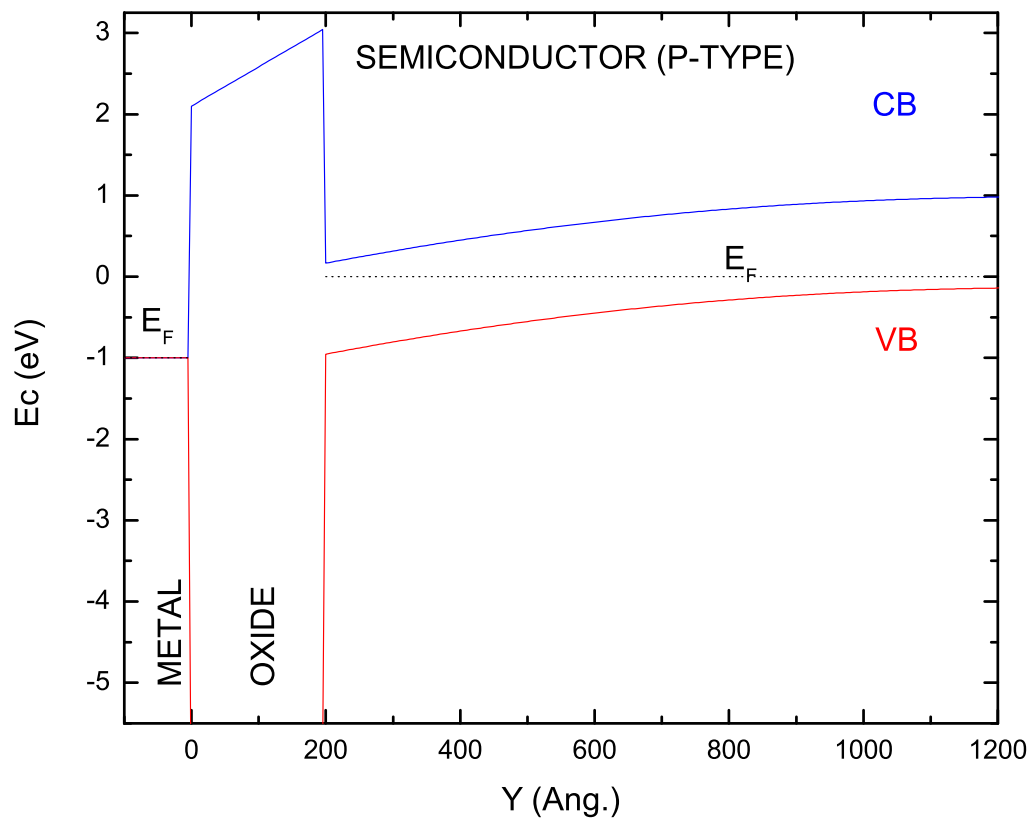


Figure 3.4: MOSFET in depletion region.

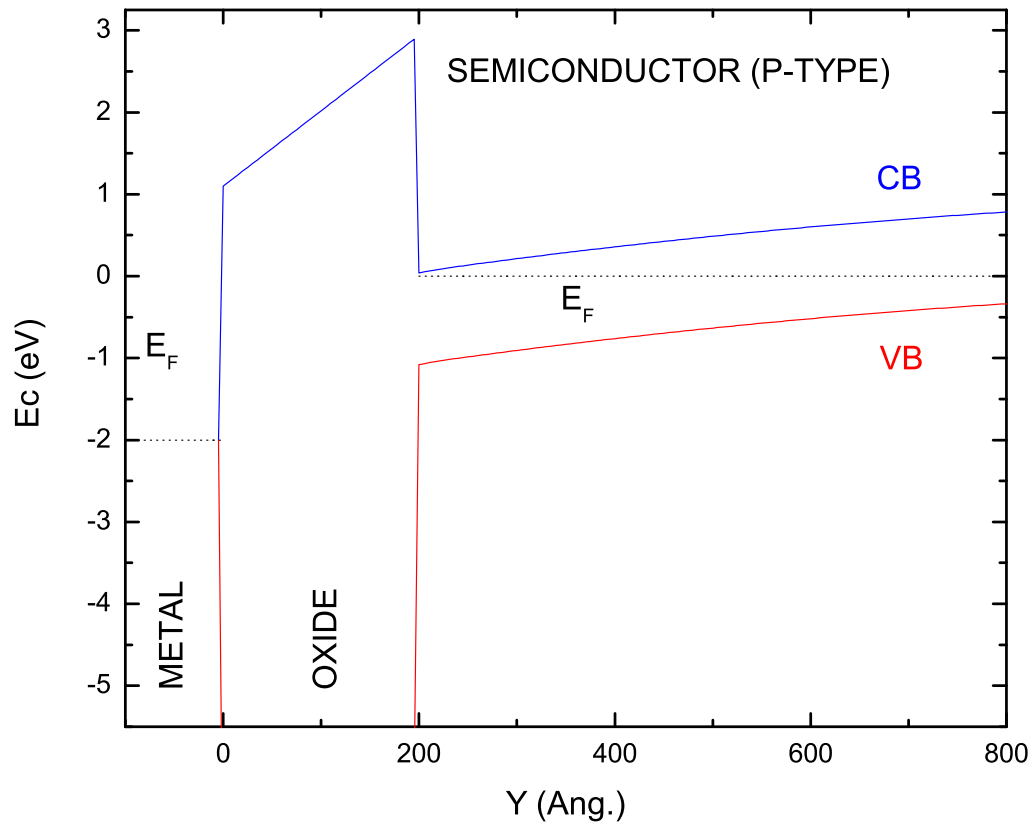


Figure 3.5: MOSFET in depletion region.

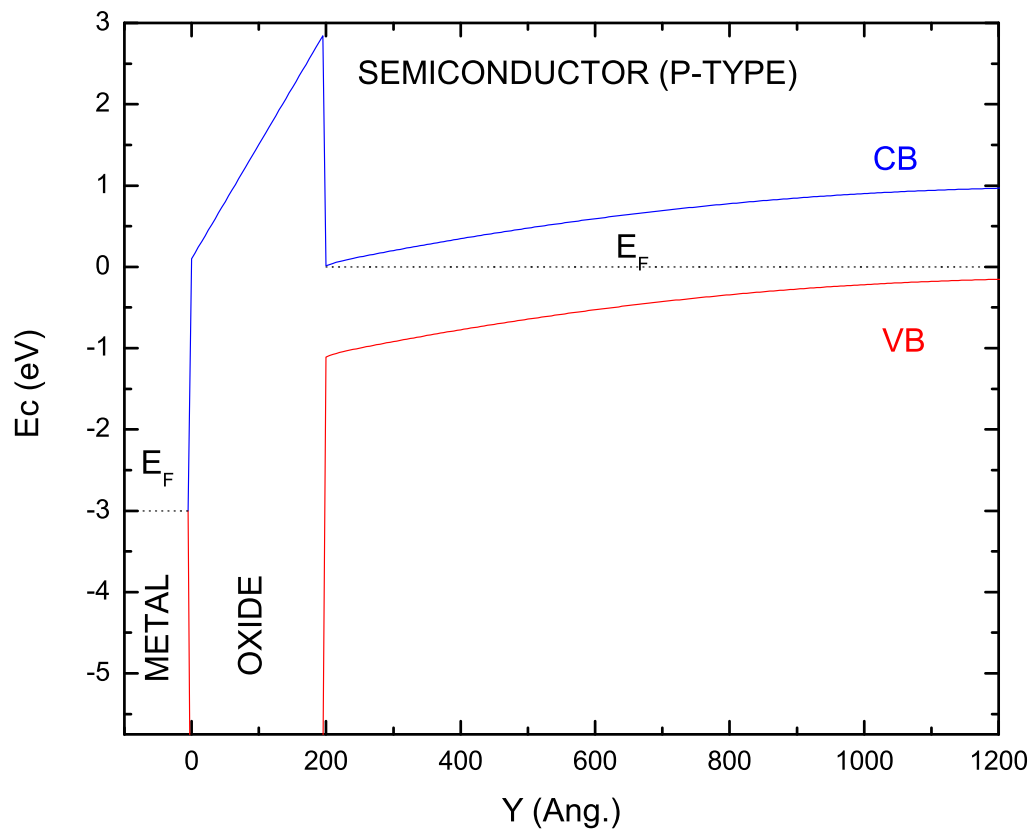


Figure 3.6: MOSFET in inversion region.

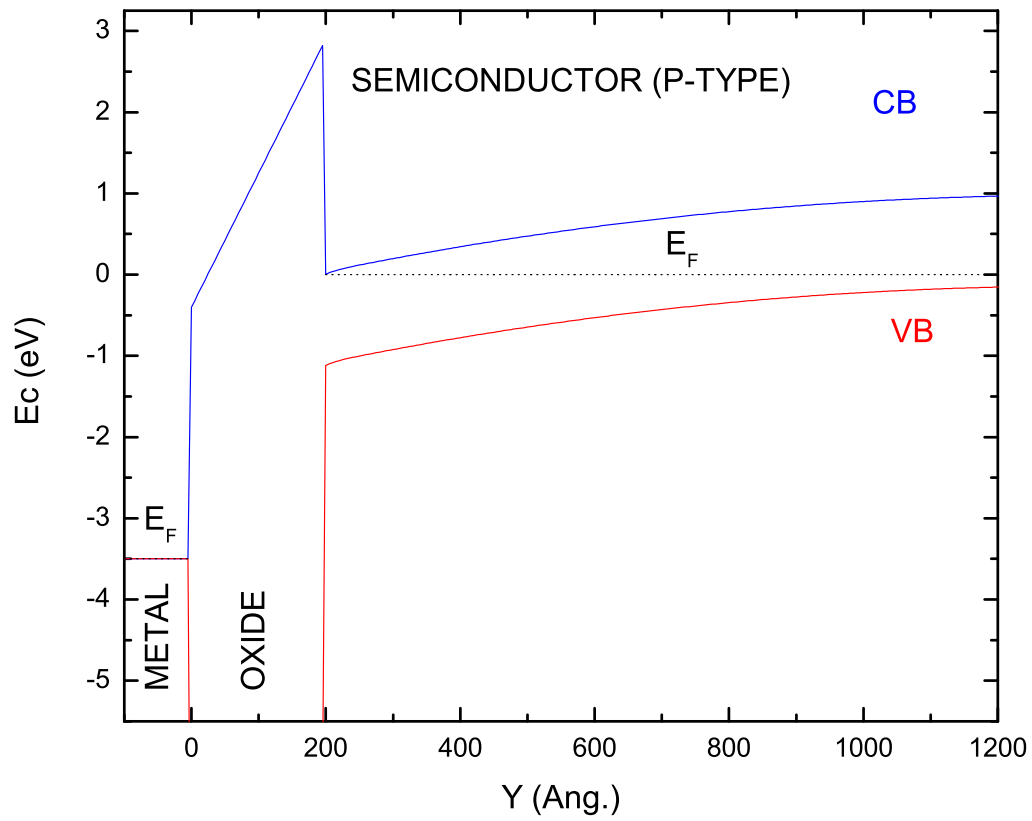


Figure 3.7: MOSFET in inversion region.

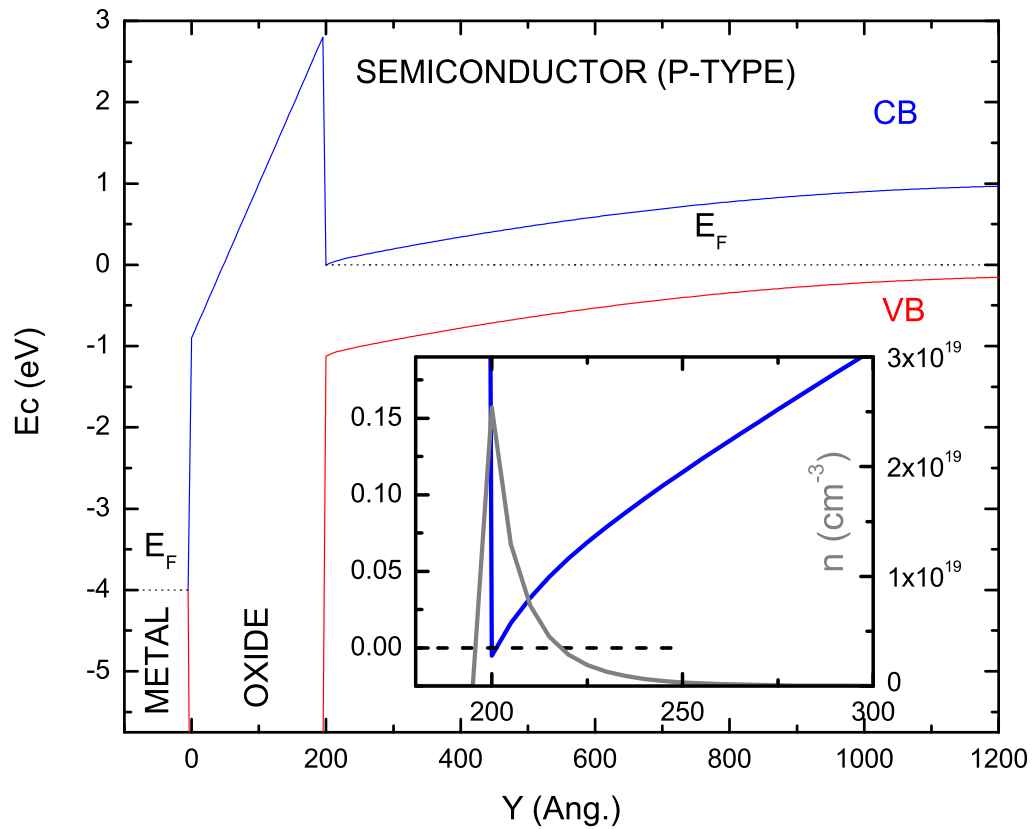


Figure 3.8: MOSFET in inversion region.

3.1.1 At what voltage will the MOSFET's channel start forming ?

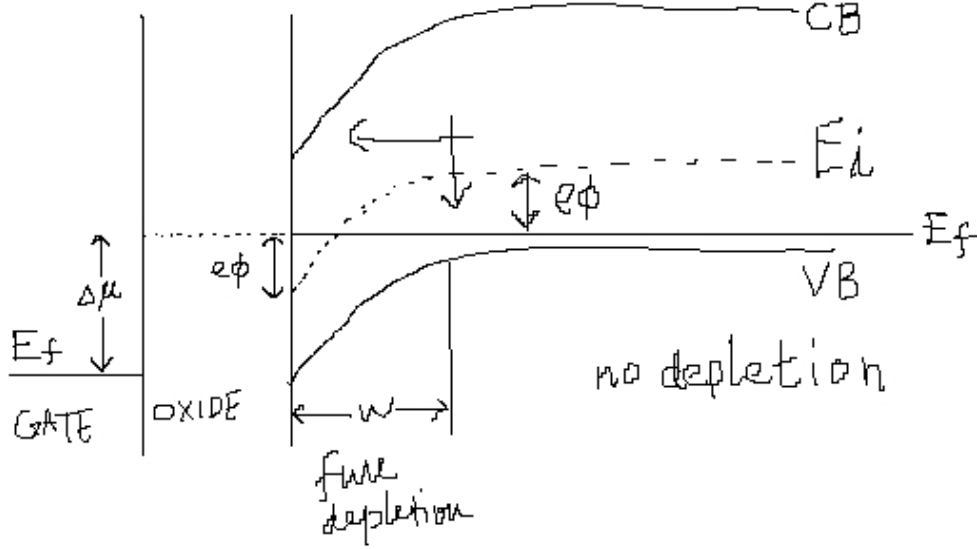


Figure 3.9: MOSFET bands where the intrinsic level dips below the E_f by exactly the same amount by which it was above E_f in the bulk.

In Fig. 3.9 the MOSFET is biased by an amount $\Delta\mu = E_{f,\text{gate}} - E_{f,\text{bulk}}$. Since there is no current flow perpendicular to the surface we assume that E_f is constant inside. Deep inside the surface the hole density is given by :

$$p = n_i e^{\beta(E_i - E_f)} \quad (3.1)$$

If we assume that all carriers there came from the dopants ($N_a \text{ cm}^{-3}$), then we get, with reference to the Fig. 3.9

$$\phi = \frac{kT}{e} \ln \frac{N_a}{n_i} \quad (3.2)$$

We have made a simplifying assumption that the depletion zone is sharp, and after the depletion zone the bands are flat. As the bias on the gate increases the battery puts more positive charge on the gate. The corresponding negative charge must go into the Si. But there are only two places where the Si can accommodate the negative charge. It either goes to the conduction band (if it allowed by the relative location of E_C w.r.t E_f) or it can sit on some acceptor site. Initially the conduction band can't accommodate any electrons because it is far above E_f . So the negative charge injected by the battery must sit on acceptor sites. But the number of charge per unit volume that can be accommodated this way is limited by the number of acceptors ($N_a \text{ cm}^{-3}$). So as more charge comes in the depletion zone must get wider and thus making the band bending start from deeper inside the Si. This bending would be roughly parabolic, as a consequence of Poisson's eqn. It is clear that if the bending begins deep enough inside the Si then at the surface the bands may dip sufficiently so that the conduction band has appreciable probability of having electrons. This is the "inversion" region of the MOSFET, where the thin layer of electrons (the channel) is separated from the p-type bulk by the depletion region.

The depletion region width at threshold is assumed to be "w" and it contains an amount of charge $Q = -|e|N_a w$. In the depletion region, we have for the electrostatic potential V ,

$$\frac{d^2 V}{dx^2} = \frac{|e|N_a}{\epsilon_0 \epsilon_r} \quad (3.3)$$

It is convenient to set the origin temporarily to where the bending just begins (see fig 3.9). Then solving eqn 3.3 gives :

$$2\phi = \frac{|e|N_a w^2}{2\epsilon_0\epsilon_r} \quad (3.4)$$

$$= \frac{Q^2}{2\epsilon_0\epsilon_r N_a |e|} \quad (3.5)$$

Now we calculate how much energy (U) is stored in the electric field in the oxide and the Silicon , between 0 and w .

$$U = U_{oxide} + U_{Si} \quad (3.6)$$

$$= \frac{Q^2}{2C_{ox}} + \frac{\epsilon_0\epsilon_r}{2} \int_0^w dx E^2 \quad (3.7)$$

$$= \frac{Q^2}{2C_{ox}} + \frac{Q^3}{6\epsilon_0\epsilon_r N_a} \quad (3.8)$$

This energy must be a consequence of the work done by the battery in creating the electrochemical potential difference $\Delta\mu$ between the gate and the Si substrate. Hence we must have

$$\int_0^{Q_{th}} \frac{\mu}{e} dQ = \frac{Q^2}{2C_{ox}} + \frac{Q^3}{6\epsilon_0\epsilon_r N_a} \quad (3.9)$$

Differentiating w.r.t Q_{th} , we get

$$\frac{\Delta\mu(Q_{th})}{e} = \frac{Q_{th}}{C_{ox}} + \frac{Q_{th}^2}{2\epsilon_0\epsilon_r N_a |e|} \quad (3.10)$$

$$= \frac{(4\epsilon_0\epsilon_r N_a |e| \phi)^{1/2}}{C_{ox}} + \frac{2kT}{e} \ln \frac{N_a}{n_i} \quad (3.11)$$

In the last step we have used eqn 3.5. We implicitly assumed that at the beginning, when the battery was not connected the bands were flat. This is not entirely correct and the voltage required to get to the flatband condition is usually a device parameter that depends on several design factors. The result should be interpreted as the excess voltage required after reaching the flatband condition : *i.e.* $V - V_{fb}$

3.2 The GaAs-AlGaAs heterostructure

A heterostructure is a generic name for two different semiconductors grown on top of each other. Experimentally it is different from depositing a thin film of one material on top of another. The heterostructure growth process usually involves growing one type of material in a vacuum chamber, then growing the other one *without breaking the vacuum*. The interface where the composition changes may be "atomically sharp". This is near impossible to achieve by depositing one type of material on another.

While many such combinations can be tried two of these systems stand out, because of their close lattice match. See Fig 3.10.

- The GaAs-Al_xGa_{1-x}As system over the full range of composition, a favourite for high mobility 2DEGs, 1D-channels, quantum dots.
- The In_{0.53}Ga_{0.47}As-In_{0.52}Al_{0.48}As-InP system, used for high speed electronics

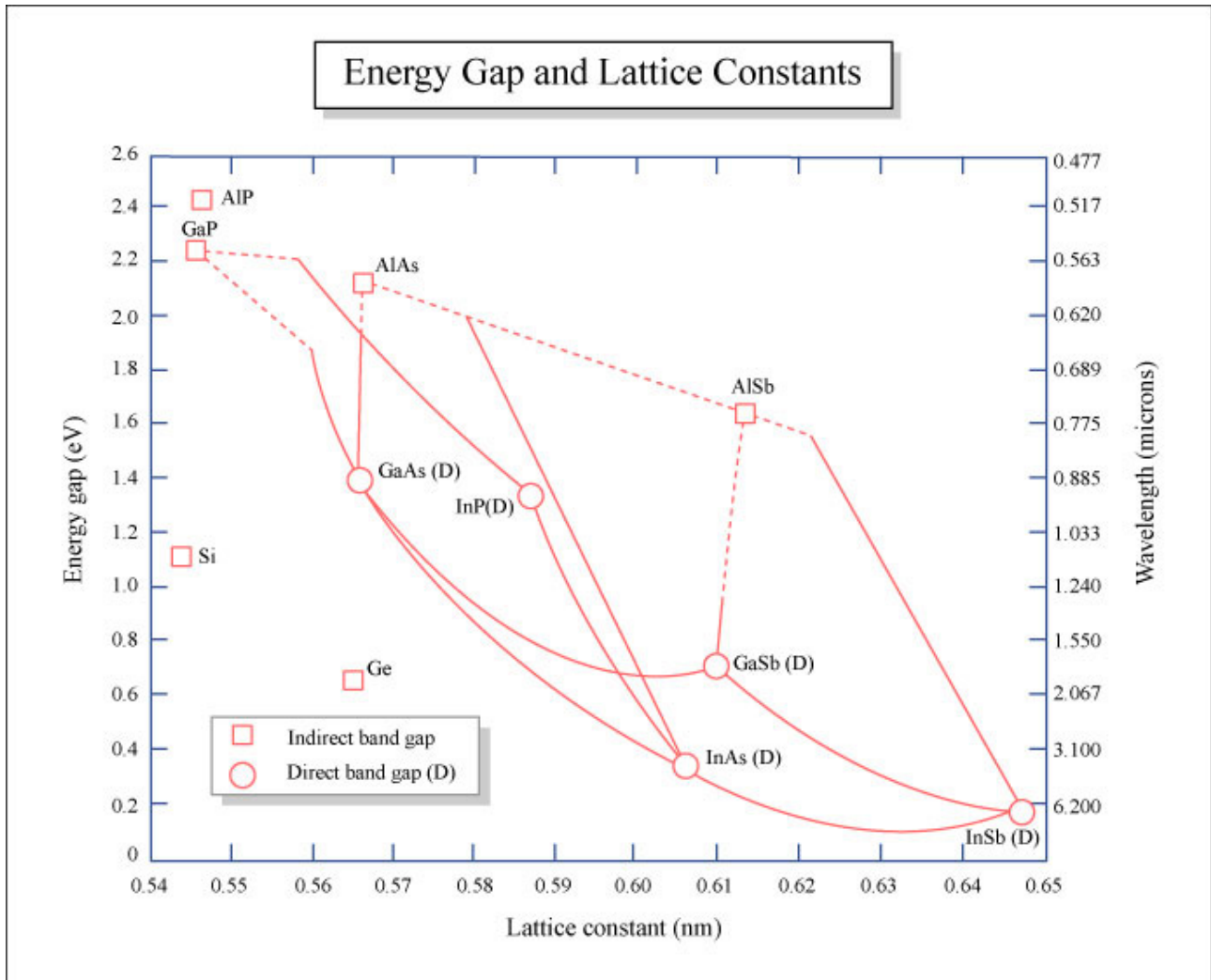


Figure 3.10: Energy gap and lattice constants of some semiconductors. To make a heterostructure with a smooth interface we must choose a pair with very close lattice matching. Otherwise the interface will have lots of strains and dislocations.

3.2.1 Anderson's rule, alignment of the bands at the heterointerface

Now, when we put two semiconductors with dissimilar band gaps side by side, an obvious question is how should the difference in the gap be distributed between the conduction and valence bands?

Without knowing this we cannot draw the band diagram, because we do not know where to place the CB of GaAs w.r.t to AlGaAs, for example, when they are in contact.

The question has no easy answer, in fact the answer needs to be found experimentally. There is however one historical attempt : Recall that the electron affinity of a semiconductor is defined as the energy needed to take an electron from the bottom of the conduction band to the vacuum level. The reason we talk about electron affinity (χ) rather than work function (Φ) is that (Φ) is measured from the Fermi level. For a semiconductor the location of the Fermi level depends on doping. Thus if we align the vacuum levels, then the obvious picture suggests that $\Delta\chi$ should translate to ΔE_C . Does this work? For GaAs-AlGaAs, we know experimentally that $\chi_{\text{GaAs}} = 4.07\text{eV}$ and $\chi_{\text{AlGaAs}} = 3.74\text{eV}$. So the $\Delta E_C = 0.33\text{eV}$ should be expected. In reality it is about 0.25 eV.

Type I, II and III band alignments

- Type 1: The smaller gap is "within" the larger gap, so that there is a barrier for both electron and hole motion if we approach from the side of the smaller gap material. *e.g* GaAs-AlGaAs
- Type 2: The smaller gap is only partially intercepted , so it is a barrier for electrons but a "drop" for holes or vice-versa. *e.g* InAs-AlSb
- Type 3: The gaps don't intercept at all. *e.g* GaSb-InAs

The concept of remote doping

The idea that the ionised dopants and the carriers may be separated in space is a very remarkable one - it allows us to increase the mobility of carriers in the channel by almost 10,000 times and the number is still increasing. We will analyse, how this increase comes about later (We need to develop simple theories of screening and Coulomb scattering before that). First we see how the band structure looks as a consequence of remote doping.

3.2.2 Band structure of the GaAs-AlGaAs heterostructure

The schottky barrier at the surface of GaAs as well as the substrate end is assumed to be 0.7 V. This means that at the two ends the conduction band of GaAs will be 0.7V above the Fermi level. Since the device is not connected to any voltage source, we must have the Fermi level to be constant through the device. As discussed earlier, since we are essentially solving differential equations, it is necessary to provide sufficient boundary conditions. The height of the barrier provides that information. Without specifying this we couldn't have solved the equations numerically. In the following band diagrams we have assumed a structure with length units in Å and doping in cm^{-3} .

surface schottky=0.7 V

GaAs t=167 Å

AlGaAs t=400 Å x=0.33 Nd= $5 \times 10^{17} \text{ cm}^{-3}$

AlGaAs t=400 Å x=0.33

GaAs t=1000 Å

GaAs t=2000 Å

substrate schottky=0.7 V

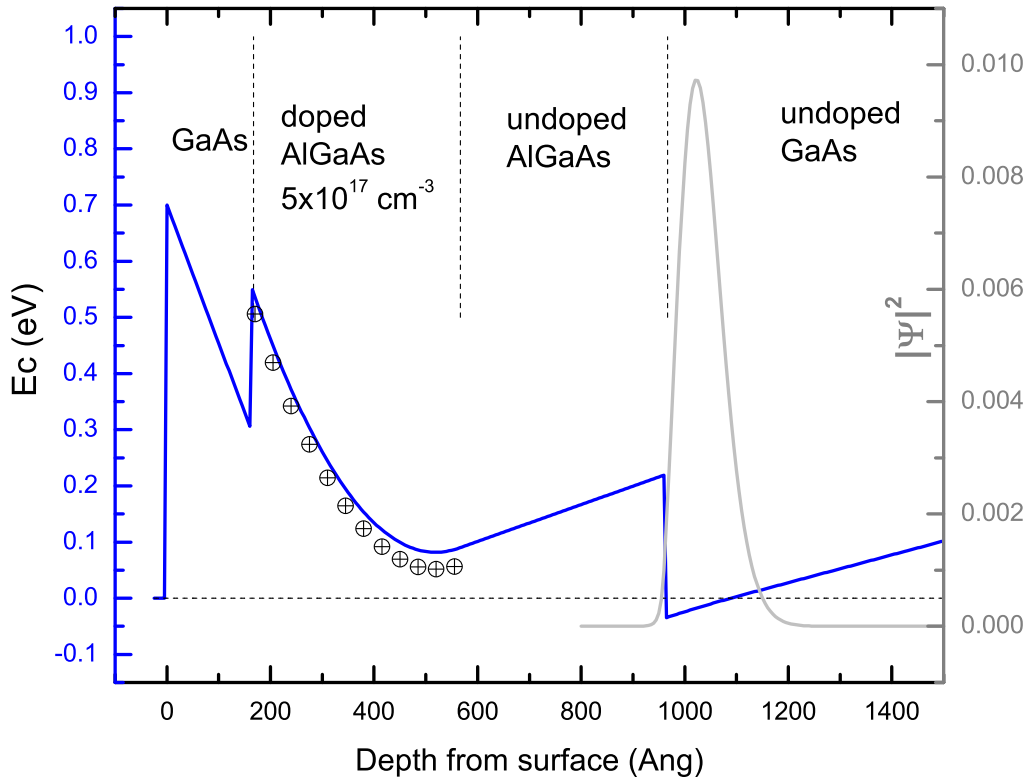


Figure 3.11: Band structure and electron wavefunction with a doping of $5 \times 10^{17} \text{ cm}^{-3}$ over 40nm and then a 40nm spacer. The AlGaAs has an Al fraction of 0.33. The dopant level is approximately 30 meV below the conduction band - it is a shallow donor.

And then increased the doping

surface schottky=0.7 V

GaAs t=167 Å

AlGaAs t=400 Å x=0.33 Nd= $1 \times 10^{18} \text{ cm}^{-3}$

AlGaAs t=400 Å x=0.33

GaAs t=1000 Å

GaAs t=2000 Å

substrate schottky=0.7 V

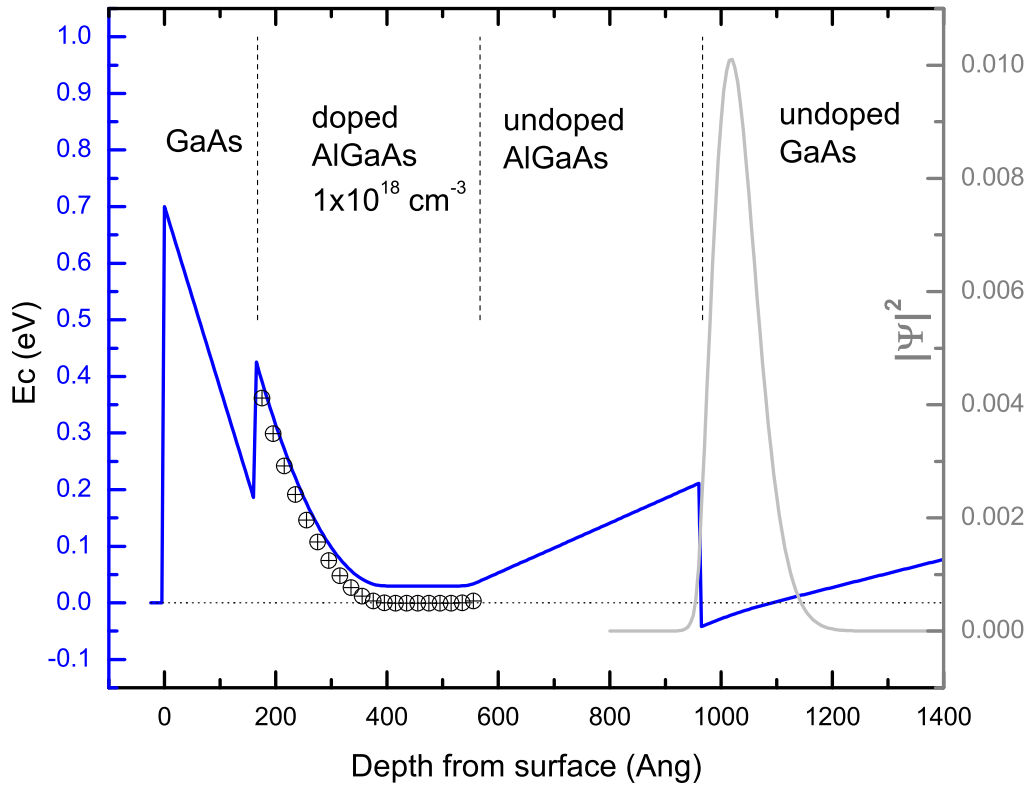


Figure 3.12: Band structure and electron wavefunction with a doping of $1 \times 10^{18} \text{ cm}^{-3}$ over 40nm and then a 40nm spacer. The AlGaAs has an Al fraction of 0.33. The dopant level is approximately 30 meV below the conduction band - it is a shallow donor.

Notice that if the dopants are to be fully ionised then the donor levels must be well above the Fermi energy and the bands in the doped region will be approximately parabolic. Similarly to a good first approximation the bands in the region between the 2D electron gas and the dopants must be linear (Poisson's equation with no charge density...). In Fig. 3.12 the doping was sufficiently high to pull the donor levels very close to the Fermi energy and you can see the "flattening" of the bands over a small region. If the doping is increased even further a pocket of electrons will start gathering there and the doped region will start behaving somewhat like a bulk conducting region. Such "parallel conducting" layers are usually undesirable in heterostructures.

PROBLEM : Sketch qualitatively how the band structure would look at threshold of electron accumulation. You only need to apply Gauss's law qualitatively to do this - you don't need a computer!

3.3 The envelope function approximation

When we solve the self consistent equations for getting the band bottom and the charge density classically it is clear that we are :

- Allowing the Fermi level to be constant - or known at every point.
- Writing the charge density at a point as a sum total of the ionised dopant charge + free electrons.

$$\begin{aligned}
 \rho(x) &= |e|N_D^+(x) - |e|n(x) \\
 N_D^+(x) &= \frac{N_D(x)}{1 + \frac{1}{2} \exp\left(\frac{E_F - E_D}{kT}\right)} \\
 n(x) &= N_c(x) \exp\left(\frac{E_F - E_C}{kT}\right) \\
 \frac{d^2 E_C}{dx^2} &= -\frac{|e|\rho(x)}{\epsilon_0 \epsilon_r}
 \end{aligned} \tag{3.12}$$

- This set of equation would give us the classical solution for the charge density. The only quantum mechanical bit that enters here is the concept of a Fermi level.

But looking at the band structure it becomes clear that the free electron density is confined in narrow regions. That means, quantum mechanical confinement etc should be taken into account. How should it be done.

First we must define clearly how we can introduce a wavefunction in the problem. Clearly it is a bit different from a single particle problem. Besides the full potential seen by the electrons should have parts coming from the "fast" variation of the atomic cores and a the slower part from the variation due to dopants/ defects/extra charge density/gates etc. We really need a solution to

$$\left[\frac{\mathbf{p}^2}{2m_0} + V_{lattice}(\mathbf{r}) + V_{slow}(\mathbf{r}) \right] \Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{3.13}$$

where $V(\mathbf{r})$ is the sum total of all the other potentials. We want to factor out the fast variations due to the lattice atom cores and get a function (*envelope*) that gives the slow variation. We give an outline of how to do it¹. We have not introduced any concept of an effective mass yet. But we write the full wavefunction as a product of band-edge Bloch functions (which vary rapidly) and a slowly varying function as follows:

$$\Psi(\mathbf{r}) = \sum_n \chi_n(\mathbf{r}) u_{n,k=0}(\mathbf{r}) \tag{3.14}$$

where $u_{n,k=0}(\mathbf{r})$ are the band-edge Bloch functions. For simplicity we have kept only the band index and omitted the spin index. Now:

- We substitute this in the full Schrodinger equation 3.13 and left multiply with $u_{n,k=0}^*(\mathbf{r})$

¹For full details and references see the book: "SpinOrbit Coupling Effects in Two-Dimensional Electron and Hole Systems". Roland Winkler, *Springer tracts in modern physics, vol 191 (2003)*

- Then integrate over one unit cell, assuming that $\psi_n(\mathbf{r})$ is reasonably slowly varying and can be pulled out of the integral.
- This will result in a matrix equation for all possible combinations of the band index.
- We restrict them to a few - or just one.
- A second order perturbation then gives, *with the assumption of parabolic bands*, the effective mass (m_n for the n-th band), envelope function hamiltonian:

$$\left[\frac{\mathbf{p}^2}{2m_n} + V(\mathbf{r}) \right] \psi(\mathbf{r}) = E\psi(\mathbf{r}) \quad (3.15)$$

Notice that the free electron mass has gotten replaced with the band effective mass. It is not at all obvious at first glance that would happen and the whole calculation is not very trivial either. But the final result is immensely useful. It is worth keeping in mind that it has its limits of validity. Notice that we have also ignored the presence of many electrons by ignoring electron-electron interaction. A justification of this (or its extent of validity) was analysed by Kohn².

3.3.1 A handwaving justification

Let us treat a simplified 1-d situation. We can (*without any approximation*) write $\Psi(x)$ as

$$\Psi(x) = \int_{-\frac{\pi}{a}}^{\frac{\pi}{a}} \chi(k) u_{n,k} e^{ikx} \frac{dk}{2\pi} \quad (3.16)$$

Notice that the integration runs over one Brillouin zone and hence includes all wavefunctions of the same band. The approximation will have to be carefully done. In the Hamiltonian there is a derivative operator, we can't throw away the fast bits before the derivative has acted on it. On the other hand while doing an integration that may be fine. The hamiltonian 3.13 is

$$\begin{aligned} H &= \frac{\mathbf{p}^2}{2m_0} + V_{lattice}(\mathbf{r}) + V_{slow}(\mathbf{r}) \\ &= T + V_{lattice} + V_{slow} \\ &= H_0 + V_{slow} \end{aligned} \quad (3.17)$$

Now apply H_0 on the full wavefunction 3.16.

$$\begin{aligned} H_0 \Psi(x) &= H_0 \int_{-\frac{\pi}{a}}^{\frac{\pi}{a}} \chi(k) u_{n,k} e^{ikx} \frac{dk}{2\pi} \\ &= \int_{-\frac{\pi}{a}}^{\frac{\pi}{a}} \chi(k) \varepsilon_{n,k} u_{n,k} e^{ikx} \frac{dk}{2\pi} \end{aligned} \quad (3.18)$$

$$\approx u_{n,k=0} \int_{-\frac{\pi}{a}}^{\frac{\pi}{a}} \chi(k) \varepsilon_{n,k} e^{ikx} \frac{dk}{2\pi} \quad (3.19)$$

Now note:

- The approximation utilised the fact that near the band edge, the Bloch-functions don't change very fast and we can write $u_{n,k} \approx u_{n0}$ and that $\chi(x)$ does not contain very fast (large k) components. The two assumptions are very crucial but not inconsistent.

²W. Kohn, *Physical Review*, **105**(2), 509516 (1957)

- ε_{nk} is a polynomial in k .
- What happens when we multiply the fourier transform of a function (in this case $\chi(k)$) with a power of k and take the inverse transform?
- If the concerned power is k^n then is same as taking the n^{th} derivative of the function and then taking its transform.

Hence the last equation can be written fully in real-space (quite strikingly!) as

$$H_0 \Psi(x) = u_{n,k=0} \varepsilon_n(-i \frac{d}{dx}) \chi(x) \quad (3.20)$$

where $\varepsilon_n(-i \frac{d}{dx})$ means that the variable k has been substituted by the momentum operator (except for the \hbar). Since near the band extremum we can approximate the dispersion by a second order term (k^2) in the effective mass approximation, we can see the origin of the effective mass term in eqn. 3.15.

The action of V_{slow} on $\Psi(x)$ is easy to approximate

$$\begin{aligned} V_{slow} \Psi(x) &= V_{slow} \int_{-\frac{\pi}{a}}^{\frac{\pi}{a}} \chi(k) u_{n,k} e^{ikx} \frac{dk}{2\pi} \\ &\approx V_{slow} u_{n,0}(x) \int_{-\frac{\pi}{a}}^{\frac{\pi}{a}} \chi(k) e^{ikx} \frac{dk}{2\pi} \\ &= V_{slow} u_{n,0} \chi(x) \end{aligned} \quad (3.21)$$

and so the action of the hamiltonian 3.17 now reduces to

$$\begin{aligned} u_{n,0}(x) \varepsilon(-i \frac{d}{dx}) \chi(x) + u_{n,0}(x) V_{slow}(x) \chi(x) &= u_{n,0}(x) E \chi(x) \\ \left[\frac{p^2}{2m_{eff}} + V_{slow} \right] \chi(x) &= E \chi(x) \end{aligned} \quad (3.22)$$

By keeping terms upto second order in the expansion of $\varepsilon(-i \frac{d}{dx})$ the EFA result follows, because $u_{n,0}$ cancels from each side. It is striking how the form of the Schrodinger equation is retained.

3.3.2 Position dependent effective mass

Now we will state without proof two important results about what happens near the boundary of two semiconductors. Without this we can't get useful results from a differential equation! Also important is the question, what happens when m_{eff} varies with position. This will happen if the composition of the semiconductor changes.

first, the correct boundary conditions (at an interface A-B assumed to be at $x = 0$) are

$$\chi_A(0) = \chi_B(0) \quad (3.23)$$

$$\frac{1}{m_A} \frac{d\chi_A}{dx} = \frac{1}{m_B} \frac{d\chi_B}{dx} \quad (3.24)$$

Infact the envelope function gets a kink at an interface if $m_A \neq m_B$

Second, the correct way to write the EFA "Schrodinger" equation, *so that it is in the "Sturm-Liouville" form*, is to include the effective mass within the derivative. This would ensure that the equation has real eigenvalues etc.

$$-\frac{\hbar^2}{2m_0} \frac{d}{dx} \left[\frac{1}{m(x)} \frac{d\chi}{dx} \right] + V_{slow} \chi(x) = E \chi(x) \quad (3.25)$$

All the "wavefunctions" are normalised to unity over the allowed interval. We denote each eigenvalue by E_j . Now we can write down the set of equations that can be solved to get a quantum version of the bands and charge densities at the interface in place of the set of equations 3.12:

$$\begin{aligned}
\rho(x) &= |e|N_D^+(x) - |e|n(x) \\
N_D^+(x) &= \frac{N_D(x)}{1 + \frac{1}{2} \exp(\frac{E_f - E_D}{kT})} \\
n(x) &= \sum_j |\chi_j(x)|^2 \frac{m^*}{\pi \hbar^2} \int_{E_j}^{\infty} \frac{dE}{1 + \exp \frac{E - E_f}{kT}} \\
\frac{d^2 E_C}{dx^2} &= -\frac{|e|\rho(x)}{\epsilon_0 \epsilon_r}
\end{aligned} \tag{3.26}$$

- There is a possibility of more than one subbands being occupied if either the temperature or the electron density is high enough.
- We need to put in the discontinuity of $E_C(x)$ at the heterointerface wherever needed.

3.4 Using the envelope function : Fang Howard wavefunction at a heterointerface

Most of the times interfacial wavefunctions are solved numerically. Self consistent schemes however don't yield analytic results in general. There are times when an analytic expression helps - even if it is a bit approximate. The Fang-Howard wavefunction is such a case, it is also an example where the interaction between electrons are taken care to an extent. This was originally developed for MOS-FETs but is applicable to almost any heterointerface where a 2DEG accumulates, with some limitations.

We know that the potential profile is roughly triangular and so a generic shape of χ can be guessed, which rises and then falls roughly as shown. We will try a variational approach, including electron-electron interaction.

$$\chi(z) = \left(\frac{b^3}{2}\right)^{1/2} z e^{-bz/2} \tag{3.27}$$

Hence the charge density is given by:

$$\rho(z) = -en_{2D}|\chi(z)|^2 = -en_{2D} \frac{b^3}{2} z^2 e^{-bz} \tag{3.28}$$

The function has a single variational parameter and the b^3 is necessary on dimensional grounds (can you see why?) The potential energy (called the Hartree potential energy here) must be

$$V_H(z) = -e\phi_H(z) \tag{3.29}$$

where

$$\begin{aligned}
\frac{d^2 \phi_H(z)}{dz^2} &= -\frac{\rho(z)}{\epsilon_0 \epsilon_r} \\
&= -\frac{en_{2D}}{\epsilon_0 \epsilon_r} \frac{b^3}{2} z^2 e^{-bz} \\
\therefore \phi_H(z) &= -\frac{en_{2D}}{\epsilon_0 \epsilon_r} \left[6 - (b^2 z^2 + 4bz + 6)e^{-bz} \right]
\end{aligned} \tag{3.30}$$

where we have used the boundary conditions

$$\frac{d\phi}{dz} = 0 \quad \text{as } z \rightarrow \infty \quad (3.31)$$

$$\phi(z=0) = 0 \quad (3.32)$$

Now

- There is no external potential in the system. So The hamiltonian for each electron is simply $H = T + V_H$.
- So the obvious step would be to calculate the energy expectation and minimize it w.r.t the only free parameter b .
- But this is wrong. What we want to minimize is the total energy of the system. If we simply add the kinetic and potential energy parts calculated for a single particle and add it n times, we double count the potential energy. The Coulomb potential needs to be counted per pair.
- This means that we should minimise $\langle T \rangle + \langle V_H \rangle / 2$

This is left as an exercise to show that

$$\langle T \rangle = \frac{\hbar^2 b^2}{8m} \quad (3.33)$$

$$\langle V_H \rangle = \frac{33e^2 n_{2D}}{16\epsilon_0 \epsilon_r b} \quad (3.34)$$

PROBLEM : Prove the two results: by calculating the expectation values

Which leads to

$$b = \left(\frac{33me^2 n_{2D}}{8\hbar^2 \epsilon_0 \epsilon_r} \right)^{1/3} \quad (3.35)$$

The one electron energy level (the lowest subband) is given by:

$$\varepsilon_1 = \left[\frac{5}{16} \left(\frac{33}{2} \right)^{2/3} \right] \left[\frac{\hbar^2}{2m} \left(\frac{e^2 n_{2D}}{\epsilon_0 \epsilon_r} \right) \right]^{1/3} \quad (3.36)$$

- Notice that the shape of the wavefunction is strongly dependent on the electron density. if n_{2D} is large then b is large and the wavefunction is narrow. That means that at low densities the envelope is actually more spread out.
- This may be counterintuitive at first glance but is absolutely correct! This is an important point for calculating screening and scattering, as we will see later.
- One weakness of this approach is that we have assumed there is no leakage of the wavefn into the larger bandgap material. This is not entirely correct for the GaAs-AlGaAs interface becuase the barrier is not very high (about 250 meV typically). However there is no simple analytic way of adding this bit. So we often decide to live with this bit of inaccuracy for the sake of the remarkable simplicity of the Fang-Howard wavefunction.

Chapter 4

Thomas Fermi screening in the electron gas : Why does remote doping lead to high mobilities?

References:

1. Chapter 17 *Solid state physics*, N. W. Ashcroft and D. Mermin
 2. Chapter 9 *Physics of low-dimesnional semiconductors* J.H. Davies
-

We want to be able to understand why in a heterostructure much higher mobilities are possible compared to bulk GaAs or MOSFETS. For this we need to develop some background to understand scattering due to impurities and how screening due to free electrons work.

4.1 Polarisation of the lattice

All elementary electrostatics books have something about dielectrics. In a dielectric there are no free electrons *i.e* nothing in the conduction band (CB). Each lattice site has atoms with bound electrons. When an external electric field (or potential) is applied, the electron wavefunction around the atom shifts a little bit. The result is that (in most cases) each atom becomes like a small dipole. The total dipole moment per unit volume is related to the polarisation (\mathbf{P}) as

$$\nabla \cdot \mathbf{P} = -\rho_{induced} \quad (4.1)$$

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E} \quad (4.2)$$

If there are free charges then something more must happen, because the free charges in the CB would reorganise themselves more readily in response to a potential. We are going to consider a situation where the external potential is due to some fixed charges - like impurities or surface gates. Thus the external field does not vary with time, it is a "static" response that we want to calculate.

4.1.1 Jellium

The bulk material is electrically neutral, because the charge of the free electron gas is exactly cancelled by the positively charged lattice. To simplify things we totally ignore the crystal structure or discreteness of the lattice at this point. We assume that there is an uniform background (like a block of jelly) in which the electrons move. The equivalent positive charge is spread out in an average structureless way.

4.1.2 Screened potential

Consider now the following basic problem. An external charge ρ_{ext} is placed in a sea of electrons. As a result some charge redistribution will take place creating an additional charge density ρ_{ind} . Let us assume that initially the potential was zero everywhere. The potential must now satisfy

$$\nabla^2 V = -\frac{\rho_{ext} + \rho_{ind}}{\epsilon} \quad (4.3)$$

where $\epsilon = \epsilon_0 \epsilon_r$.

Notice that we can solve this, if we relate ρ_{ind} to something known. Recall that in equilibrium with no battery connected to the system, the Fermi level (E_f) must be constant. But if the electrostatic potential varies, then the bottom of the CB also varies, as shown in Fig 4.1. In a place where the bottom of the CB is high some electrons must be pushed out. How many?

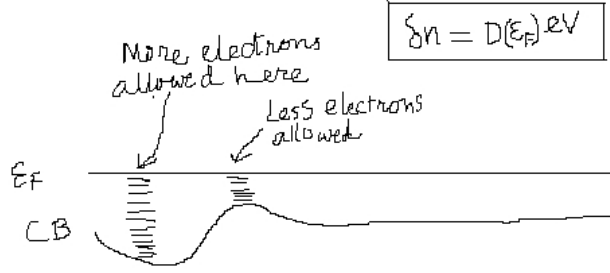


Figure 4.1: Whenever the conduction band rises (electrostatic potential is *lower*) some electrons are pushed out. If the electrostatic potential is higher, then more electrons tend to gather there. This is because the potential is defined with a positive charge in mind.

$$\delta n = D(E_f) eV \quad (4.4)$$

$$\therefore \rho_{ind} = -e^2 D(E_f) V \quad (4.5)$$

Where $D(E_f)$ is the density of states at the Fermi level. Here the implicit assumption is that $eV \ll E_f$ and that the potential varies slowly, on the scale of $1/k_F$

We then write eqn 4.3 in a solvable form

$$\nabla^2 V - \frac{e^2 D(E_f) V}{\epsilon} = -\frac{\rho_{ext}}{\epsilon} \quad (4.6)$$

$$\nabla^2 V - q_{TF}^2 V = -\frac{\rho_{ext}}{\epsilon} \quad (4.7)$$

Where we introduced the Thomas Fermi wavevector

$$q_{TF} = \left(\frac{e^2 D(E_f)}{\epsilon} \right)^{1/2} \quad (4.8)$$

Eqn 4.7 is solved by using Fourier transforms. In general the ∇^2 operator becomes multiplication by $-q^2$ in Fourier space.

$$V(\mathbf{r}) = \frac{1}{(2\pi)^3} \int d^3 \mathbf{q} \tilde{V}(\mathbf{q}) e^{-i\mathbf{q} \cdot \mathbf{r}} \quad (4.9)$$

$$\rho(\mathbf{r}) = \frac{1}{(2\pi)^3} \int d^3 \mathbf{q} \tilde{\rho}(\mathbf{q}) e^{-i\mathbf{q} \cdot \mathbf{r}} \quad (4.10)$$

Using the fourier components we now write:

$$-q^2 \tilde{V}(q) - q_{TF}^2 \tilde{V}(q) = -\frac{\tilde{\rho}_{ext}(q)}{\epsilon} \quad (4.11)$$

Hence

$$\tilde{V}(q) = \frac{\tilde{\rho}(q)}{\epsilon(q^2 + q_{TF}^2)} = \frac{\tilde{V}_{ext}(q)}{1 + \frac{q_{TF}^2}{q^2}} \quad (4.12)$$

We have thus related the external potential to the total potential and the induced charge density by eqns 4.12 and 4.5.

PROBLEM : Due to the $D(E_f)$ factor, the Thomas-Fermi wavevector depends on the dimensionality of the system. Show that in 3D :

$$q_{TF}^2 = \frac{1}{\pi^2} \frac{me^2}{\epsilon \hbar^2} k_f = \frac{4}{\pi} \frac{k_f}{a_B} \quad (4.13)$$

where a_B is the Bohr radius of the system.

Extent of validity of the Thomas Fermi approximation

The expression for screening that we deduced is valid for small q , as q approaches $2k_f$ the method no longer gives correct results.

4.1.3 The screened Coulomb potential

What will be the total potential due to a single point charge placed in a sea of electrons? For a point charge (e) at the origin $\rho(r) = e\delta(r)$. Hence the fourier transform is just a constant $\tilde{\rho}(q) = 1$. Eqn. 4.12 give

$$\tilde{V}(q) = \frac{e}{\epsilon(q^2 + q_{TF}^2)} \quad (4.14)$$

The transform can be done by using a simple contour integral and gives

$$V(r) \sim \frac{e}{r} e^{-q_{TF}r} \quad (4.15)$$

The only "pole" that will contribute to the integral is at $z = +iq_{TF}$

4.2 Screening in the 2-dimensional electron gas

The 2DEG in a MOSFET or a GaAs-AlGaAs heterostructure is *quasi* 2-dimensional only, it is really a thin layer embedded in 3D and has to be treated as such. The 2DEG does not imply a 2-dimensional electrodynamics, where Gauss's law would have to be replaced by an integration over a circle instead of a sphere!

We will calculate the 2D Fourier transform of a static point charge placed at $z = z_0$, the charge distribution (not the wavefunction) of the 2DEG is given by $f(z)$, \mathbf{r} is the in plane co-ordinate.

$$V(\mathbf{r}, z) = \frac{1}{4\pi\epsilon} \frac{e}{\sqrt{r^2 + (z - z_0)^2}} \quad (4.16)$$

Now, the important point is that it needs to be averaged over the charge distribution in the z -direction. This is needed because the electron (in the 2DEG) may be probabilistically found at any z with the probability $f(z)$. We write this averaging leads to:

$$\bar{V}(\mathbf{r}) = \frac{e}{4\pi\epsilon} \int dz \frac{1}{\sqrt{r^2 + (z - z_0)^2}} f(z) \quad (4.17)$$

$$\bar{V}(\mathbf{k}) = \frac{e}{4\pi\epsilon} \int d^2\mathbf{r} e^{i\mathbf{k}\cdot\mathbf{r}} \int dz \frac{1}{\sqrt{r^2 + (z - z_0)^2}} f(z) \quad (4.18)$$

$$\bar{V}(k) = \frac{e}{4\pi\epsilon} \int_{-\pi}^{\pi} d\theta e^{ikr \cos \theta} \int dz \int_0^{\infty} r dr \frac{1}{\sqrt{r^2 + (z - z_0)^2}} f(z) \quad (4.19)$$

$$= \frac{e}{2\epsilon} \int dz \int_0^{\infty} r dr \frac{J_0(kr)}{\sqrt{r^2 + (z - z_0)^2}} f(z) \quad (4.20)$$

The Bessel function oscillates with a decreasing amplitude and goes to zero at infinity. Since the integral ends up taking in many periods, that nearly cancel each other, the final result ends up being quite small - a bit striking at first glance. The result, using a standard integral (Gradshteyn & Ryzhik 6.554.1):

$$\int_0^{\infty} dr \frac{r J_0(kr)}{\sqrt{r^2 + (z - z_0)^2}} = \frac{e^{-k|z - z_0|}}{k} \quad (4.21)$$

So we get the unscreened potential

$$\bar{V}_{uns}(k) = \frac{e}{2\epsilon k} \int dz e^{-k|z - z_0|} f(z) \quad (4.22)$$

Let's define this quantity as

$$\bar{V}(\mathbf{k}, z_0) \equiv \frac{e}{2\epsilon k} \int dz e^{-k|z - z_0|} f(z) \quad (4.23)$$

This gives the Fourier components of the unscreened potential. If, the 2DEG is of zero thickness, then

$$f(z) = \delta(z) \quad (4.24)$$

$$\bar{V}(k) = \frac{e}{2\epsilon k} e^{-k|z_0|} \quad (4.25)$$

Now, we already notice a striking fact that the unscreened potential itself decreases exponentially as it is moved away from the plane of the 2DEG. We can already have a feeling that "remote doping" dramatically reduces the strength of the ionised impurity scattering.

PROBLEM : In a Gallium Arsenide quantum well, the mobility of the 2-dimensional electron gas is found to be $30 \times 10^6 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ at a density of $2 \times 10^{11} \text{cm}^{-2}$. Given that the effective mass of these electrons is $0.067 m_0$, What is the mean free path?

4.2.1 Finite thickness with Thomas-Fermi screening

Now consider finite thickness with screening. It is algebraically somewhat messy due to the averaging we need to do at each step - but conceptually no different from what we have done earlier.

$$\nabla^2 V(\mathbf{r}, z) = -\frac{1}{\epsilon} (\rho_{ext} + \rho_{ind}) \quad (4.26)$$

The induced charge density is written as

$$\rho_{ind}(\mathbf{r}) = -e[N(\bar{V}(\mathbf{r})) - N(0)]f(z) \quad (4.27)$$

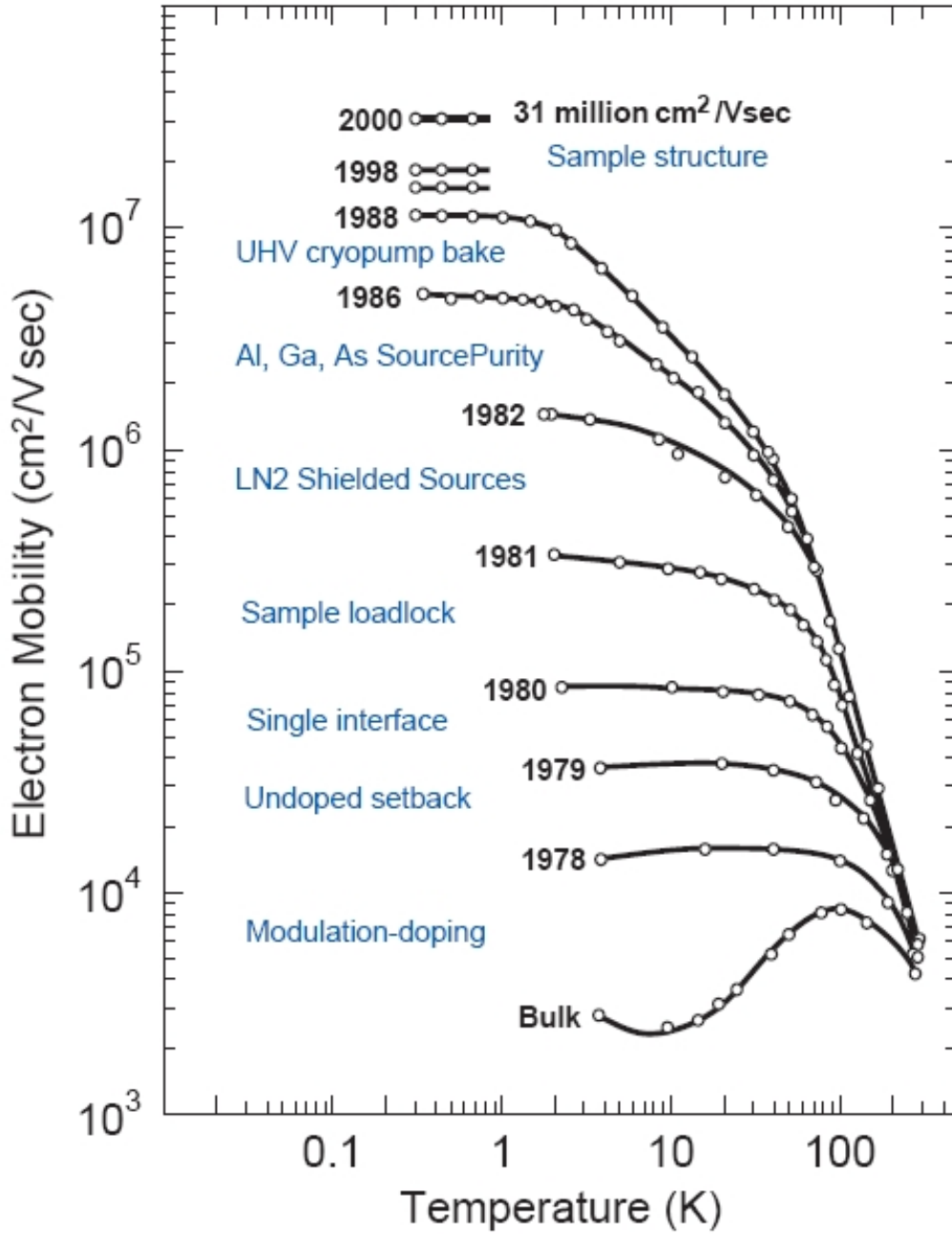


Figure 4.2: Notice the huge improvement in mobility over the years. The figure is taken from the paper by L.N. Pfeiffer, *Physica E* **20**, 57 (2003). The mobility of electrons in pure metals like Gold, Silver is much lower than that in heterostructures. A pure metal film would have higher conductivity, because it has a much much higher density of electrons.

Thomas-Fermi approximation gives

$$\rho_{ind}(\mathbf{r}, z) = -e^2 \bar{V}(\mathbf{r}) \frac{dN}{dE_F} f(z) \quad (4.28)$$

Now we define the Thomas Fermi wavevector (2D)

$$q_{TF} = \frac{e^2}{2\epsilon} \frac{dN}{dE_F} \quad (4.29)$$

$$\rho_{ind}(\mathbf{r}) = -2\epsilon q_{TF} \bar{V}(\mathbf{r}) f(z) \quad (4.30)$$

Here $\bar{V}(\mathbf{r})$ denotes the full potential at that location, not the external potential only. The equation we now need to solve is

$$\nabla^2 V(\mathbf{r}, z) - 2q_{TF} \bar{V}(\mathbf{r}) f(z) = -\frac{\rho_{ext}}{\epsilon} \quad (4.31)$$

Assume that the dielectric constant at the GaAs-AlGaAs interface doesn't change ($\epsilon_r \approx 13$). This is not correct for a Si-SiO₂ interface. Si has a much larger ϵ_r . So the analytic relations for a MOSFET would look more complicated.

We can write the equation in terms of Fourier components (notice how we handle the z part of the Laplacian)

$$V(\mathbf{r}, z) = \frac{1}{(2\pi)^2} \int d^2 \mathbf{k} V(\mathbf{k}, z) e^{-i\mathbf{k} \cdot \mathbf{r}} \quad (4.32)$$

$$\nabla^2 V(\mathbf{r}, z) = \frac{1}{(2\pi)^2} \int d^2 \mathbf{k} \left(\frac{d^2}{dz^2} - k^2 \right) V(\mathbf{k}, z) e^{-i\mathbf{k} \cdot \mathbf{r}} \quad (4.33)$$

$$\bar{V}(\mathbf{r}) = \int dz' f(z') \frac{1}{(2\pi)^2} \int d^2 \mathbf{k} V(\mathbf{k}, z') e^{-i\mathbf{k} \cdot \mathbf{r}} \quad (4.34)$$

$$\begin{aligned} \rho_{ext}(\mathbf{r}) &= e \delta^2(\mathbf{r}) \delta(z - z_0) \\ &= \frac{e}{(2\pi)^2} \int d^2 \mathbf{k} e^{-i\mathbf{k} \cdot \mathbf{r}} \delta(z - z_0) \end{aligned} \quad (4.35)$$

So we get:

$$\left(\frac{d^2}{dz^2} - k^2 \right) V(\mathbf{k}, z) - 2q_{TF} f(z) \int dz' f(z') V(\mathbf{k}, z') = -\frac{e}{\epsilon} \delta(z - z_0) \quad (4.36)$$

$$\left(\frac{d^2}{dz^2} - k^2 \right) V(\mathbf{k}, z) = 2q_{TF} f(z) \int dz' f(z') V(\mathbf{k}, z') - \frac{e}{\epsilon} \delta(z - z_0) \quad (4.37)$$

To get an explicit solution, we need to use the "Green's function" technique. We first solve (use Fourier transform and then the back transform requires a contour integral as before)

$$\left(\frac{d^2}{dz^2} - k^2 \right) G(z - z'') = \delta(z - z'') \quad (4.38)$$

The solution is

$$G(z - z'') = -\frac{1}{2k} e^{-k|z - z''|} \quad (4.39)$$

and hence

$$V(\mathbf{k}, z) = \int dz'' G(z - z'') \left[2q_{TF} f(z'') \int dz' f(z') V(\mathbf{k}, z') - \frac{e}{\epsilon} \delta(z'' - z_0) \right] \quad (4.40)$$

$$V(\mathbf{k}, z) = 2q_{TF} \int dz'' \int dz' G(z - z'') f(z') f(z'') V(\mathbf{k}, z') - \frac{e}{\epsilon} G(z - z_0) \quad (4.41)$$

But we need the value of $V(\mathbf{k}, z)$ averaged over $f(z)$, which requires another integration after multiplying both sides by $f(z)$.

$$\begin{aligned} \int dz f(z) V(\mathbf{k}, z) &= 2q_{TF} \int dz f(z) \int dz'' \int dz' G(z - z'') f(z') f(z'') V(\mathbf{k}, z') \\ &\quad - \frac{e}{\epsilon} \int dz f(z) G(z - z_0) \end{aligned} \quad (4.42)$$

The integral on z' can be pulled out, so that

$$\begin{aligned}\bar{V}(\mathbf{k}) &= 2q_{TF} \int dz \int dz'' f(z)G(z-z'')f(z'') \int dz' f(z')V(\mathbf{k}, z') \\ &\quad - \frac{e}{\epsilon} \int dz f(z)G(z-z_0)\end{aligned}\tag{4.43}$$

$$= -\frac{q_{TF}}{k} \int dz \int dz'' f(z)e^{-k|z-z''|}f(z'')\bar{V}(\mathbf{k}) - \frac{e}{2\epsilon k} \int dz f(z)e^{-k|z-z_0|}\tag{4.44}$$

Where we have used the definition of the Green's function from eqn 4.39

$$\bar{V}(\mathbf{k}) = \frac{\bar{V}_{uns}(\mathbf{k})}{1 + (q_{TF}/k)F(k)}\tag{4.45}$$

where the Form factor and the Thomas-Fermi dielectric function is defined as

$$F(k) = \int dz \int dz'' f(z)e^{-k|z-z''|}f(z'')\tag{4.46}$$

$$\epsilon(q) = 1 + \frac{q_{TF}}{k}F(k)\tag{4.47}$$

The -ve sign of $G(z-z'')$ has been absorbed in the definition of the Form factor, such that $F(k) = 1$ for $f(z) = \delta(z)$

$$(4.48)$$

Chapter 5

Quantum Hall effect : Basic physics and the edge state picture

References:

1. Chapter on Quantum Hall Effect *Physics of low-dimesnional semiconductors* J.H. Davies
 2. *The Quantum Hall Effect: Novel Excitations and Broken Symmetries*, S.M. Girvin, arXiv:condmat-mat 9907002 [Les Houches Lecture notes 1998]
-

5.1 Current flow in a rectangular "Hall Bar", the classical solution

Consider a rectangular bar of material (as shown in Fig. 5.1) that contains a 2-dimensional electron gas. There are two ohmic contacts at the two ends that we use to pass current through it. We will solve for the current (and electric field) in the rectangle, with and without a perpendicular magnetic field. The rectangle is in the x-y plane and the magnetic field (B) is uniform and along z axis only.

5.1.1 $B=0$

The hall bar has a length L and width W . We apply a potential V_0 as shown then we want to solve for the potential distribution (and electric field) inside the Hall bar.

We have

$$\nabla \cdot \mathbf{j} = 0 \quad (5.1)$$

$$\mathbf{j} = \sigma_0 \mathbf{E} \quad (5.2)$$

$$\mathbf{E} = -\nabla V \quad (5.3)$$

$$\therefore \nabla^2 V = 0 \quad (5.4)$$

What are the boundary conditions?

$$V(x, 0) = 0 \quad (5.5)$$

$$V(x, L) = V_0 \quad (5.6)$$

$$j_x(-W/2, y) = 0 \quad (5.7)$$

$$j_x(W/2, y) = 0 \quad (5.8)$$

Also the electric field is zero in the metal contacts (*i.e* the contacts are equipotential metal surfaces), as that is assumed to have a much higher conductivity than the 2DEG layer.

The solution is obvious :

$$V(x, y) = \frac{V_0}{L} y \quad (5.9)$$

5.1.2 After a magnetic field is applied

We know that the relation between \mathbf{j} and \mathbf{E} in a magnetic field ($\mathbf{B} = B_0 \hat{\mathbf{z}}$) is given by:

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = \frac{\sigma_0}{1 + \mu^2 B_0^2} \begin{pmatrix} 1 & -\mu B_0 \\ \mu B_0 & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad (5.10)$$

or

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} \rho_0 & \frac{B_0}{nq} \\ -\frac{B_0}{nq} & \rho_0 \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} \quad (5.11)$$

Clearly \mathbf{j} and \mathbf{E} are no longer parallel to each other. What is the angle (δ) between them?

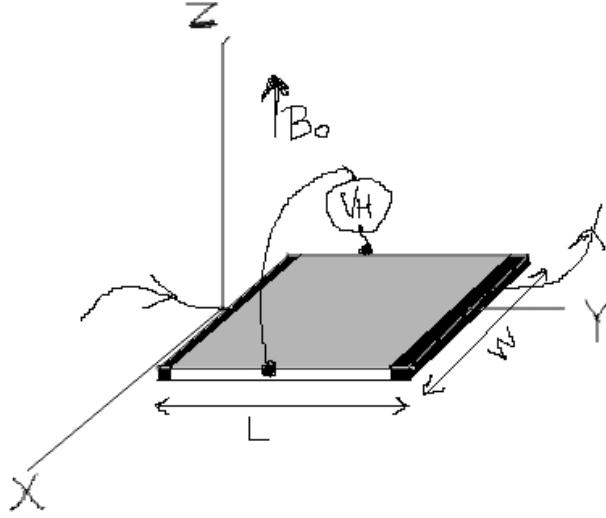


Figure 5.1: The ohmic contacts keep two ends parallel to the x-axis at the same potential.

PROBLEM : Show (using 5.11) that the Hall angle is given by

$$\tan \delta = \frac{B_0/nq}{\rho_0} \quad (5.12)$$

This is true at all points. On the edges, where ($j_x=0$) must hold, since no current is leaking out of the Hall bar, the current must point only in the y direction and hence the \mathbf{E} must make an angle δ with the y axis. As the magnetic field is made very large, this angle should tend to $\pi/2$. Unfortunately the solution can no longer be written down so easily. An analytic solution can be written in terms of $Z = x + iy$, as

$$E_y(x, y) + iE_x(x, y) = \exp(f(Z)) \quad (5.13)$$

$$f(Z) = \sum_{n(\text{odd})} \frac{4\delta}{n\pi} \frac{\sinh\left(\frac{n\pi Z}{L}\right)}{\cosh\left(\frac{n\pi W}{2L}\right)} \quad (5.14)$$

The solution and the method to obtain it is given by R.W. Rendell and S.M. Girvin, *Physical Review B*, **23**, 6610 (1981). It is done using a method of solving the Laplace's equation using a complex transform called a "conformal map", that we don't discuss here¹.

¹Most texts on Mathematical Physics, has a chapter on this method, see for example Pipes and Harvill.

Just outside the contacts the electric field only has E_y component (since the electric field must be perpendicular to the equipotential metal), but $E_y = 0$ on the sides when $\delta \approx \pi/2$. This forces a jump at two corners (see Fig. 5.2) or a mathematical singularity of the solutions. Rather than analysing the series summation, we plot out the current density and infer some physically important aspects.

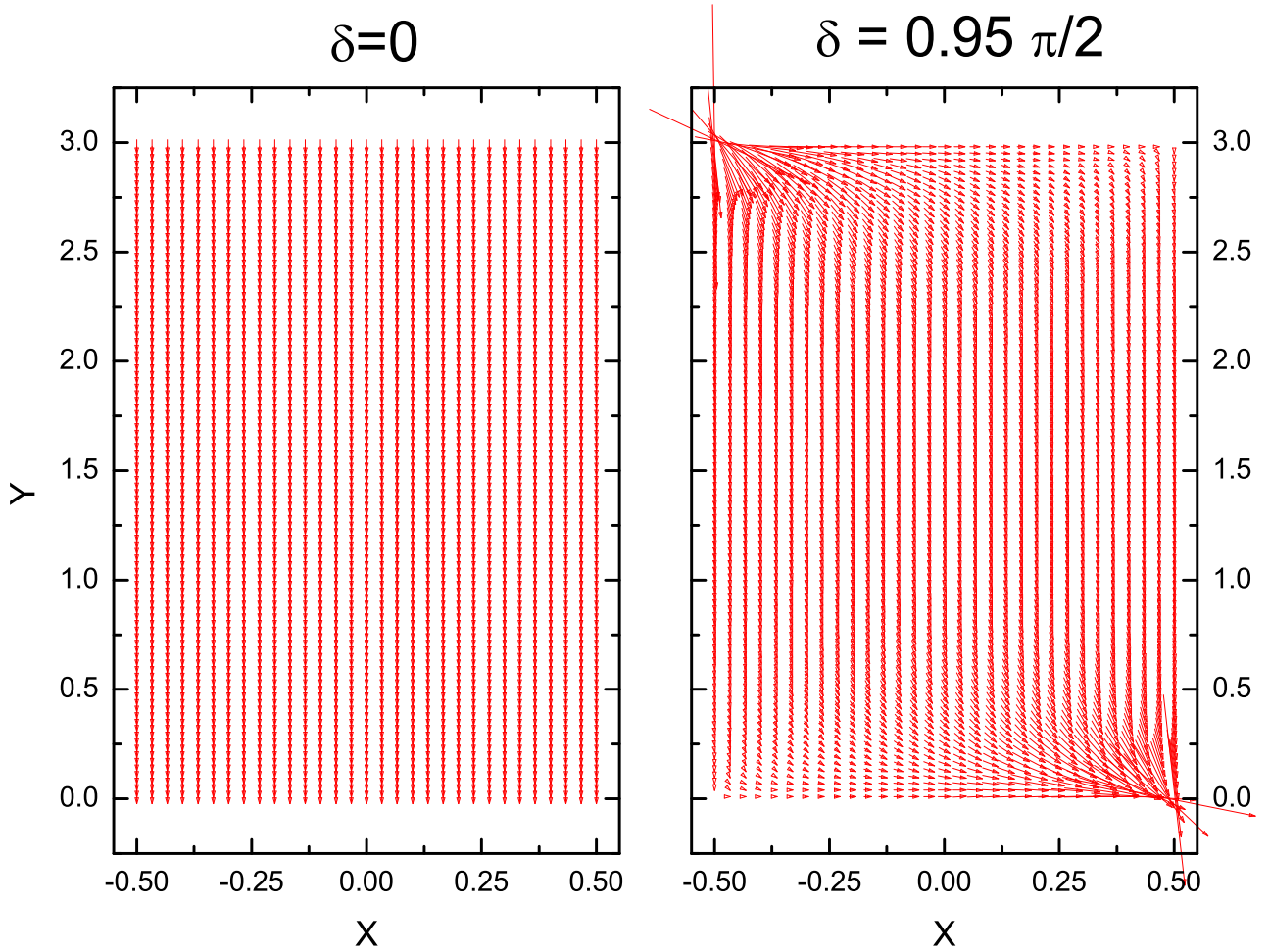


Figure 5.2: Each vector is proportional to the value of \mathbf{j} at that point. The dimensions were taken to be $L = 3$ and $W = 1$. The electric field is aligned along \mathbf{j} when $\delta = 0$, but is almost perpendicular to \mathbf{j} when δ approaches $\pi/2$.

5.1.3 Current flow directions and the equipotentials

There are several consequences of this current distribution that may be quite counter-intuitive.

- In Fig 5.2 the plot with $\delta = 0$ is easily understandable. The equipotentials must be parallel to the x -axis. The electric field is only along y and so the current (only j_y exists) is perpendicular to the equipotentials.
- But the situation is dramatically different when $\delta \approx \pi/2$. The electric field is now perpendicular to the current flow. *Hence the current flows along equipotentials.* A consequence of this is that there is no voltage drop as one follows the current!
- Thus two voltage probes placed at two points on the (long) side may measure no voltage drop.
- The current density at the corner is very high.
- All the current emanates from the edge of the ohmic contacts (the points of mathematical singularity of the series solution) and carries with it the potential of the contact.

- We did not require much "quantum" physics to establish the important fact that almost dissipationless channels can arise in a strong magnetic field, irrespective of the amount of disorder initially present. The zero field resistance ρ_0 dropped out of our consideration.
- The two-probe resistance measured between the ohmic-contacts and the Hall voltage measured between the opposite sides of the Hall bar will be the same.
- The Hall voltage measurement does not require that the Hall voltage probes be exactly opposite to each other.

5.1.4 What is *not* correct in the classical solution?

The classical solution predicts that the current is distributed over a wide area of the sample. This is not correct. In reality the states slightly far (~ 100 nm or so) from the edge are not current carrying states when the Hall angle reaches close to $\pi/2$. We need to look at the Schrodinger equation at this point.

5.2 Quantum mechanical solution

To describe a charged particle (e) in a 2D plane we can choose the following vector potential and then the Hamiltonian

$$\mathbf{A} = B_0 x \hat{\mathbf{y}} \quad (5.15)$$

$$H = \frac{1}{2m} (p_x^2 + (p_y + eB_0 x)^2) \quad (5.16)$$

Notice that in this case

$$\mathbf{p} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{p} = B_0 x p_y \quad (5.17)$$

We can now separate the variables by using the obvious solution

$$\psi(x, y) = e^{iky} f_k(x) \quad (5.18)$$

So we have the eigenvalue equation for $f_k(x)$

$$\left[\frac{p_x^2}{2m} + \frac{1}{2m} (\hbar k + eB_0 x)^2 \right] f_k(x) = E f_k(x) \quad (5.19)$$

This has the form of a harmonic oscillator equation whose center is shifted from $x = 0$. With the identifications

$$\omega = \frac{eB_0}{m} \quad (5.20)$$

$$x_0 = \frac{\hbar k}{eB_0} \quad (5.21)$$

We can write eqn 5.19 as

$$\left[\frac{p_x^2}{2m} + \frac{m\omega^2}{2} (x + x_0)^2 \right] f_k(x) = E f_k(x) \quad (5.22)$$

The solution to this equation is very well known.

$$E_n = (n + \frac{1}{2}) \hbar \omega \quad (5.23)$$

These energy levels are called the Landau levels. How much is the shift of the center of the wavefunctions?

$$x_0 = \frac{\hbar k}{eB_0} = k\ell^2 \quad (5.24)$$

$$\ell^2 = \frac{\hbar}{eB_0} \quad (5.25)$$

$$\ell \approx \frac{257\text{\AA}}{\sqrt{B}} \quad (5.26)$$

5.2.1 Degeneracy of each Landau level and the overlap of the wavefunctions

Each level has a large degeneracy, because the eigenvalue is independent of k . Consider a sample of dimension $L \times W$ as before. We must have x_0 within the sample limits, so the number of possible states N must be

$$\begin{aligned} N &= \frac{L}{2\pi} \int_0^{W/\ell^2} dk \\ &= \frac{LW}{2\pi} \frac{eB}{\hbar} \\ &= LW \frac{eB}{h} \end{aligned} \quad (5.27)$$

So the degeneracy is eB/h per unit area of the sample. The spatial width of the n^{th} harmonic oscillator wavefunctions is about $\sqrt{n}\ell$ and the separation between the centers of the wavefunctions is

$$\delta x_0 = \frac{2\pi}{L} \ell^2 \quad (5.28)$$

So for a typical sample size of $L \sim 1\text{mm}$, the states are very strongly overlapping.

5.2.2 Oscillation of E_F as B increases

Consider for the moment a spinless collection of particles - n particles per unit area in 2D. In zero magnetic field we must have

$$\begin{aligned} D(E) &= \frac{m}{2\pi\hbar^2} \\ N &= \frac{m}{2\pi\hbar^2} E_f(0) \end{aligned} \quad (5.29)$$

When a magnetic field is applied the index of the highest occupied Landau level (not the number of occupied levels, which must be at least 1) is

$$\nu_{max} = \left[\frac{N}{eB/h} \right] \quad (5.30)$$

Where the $[]$ denotes the largest integer contained in the expression. Hence the Fermi energy must be

$$E_f(B) = \left(\nu_{max} + \frac{1}{2} \right) \hbar\omega \quad (5.31)$$

$$= \left(\frac{1}{2} + \left[\frac{N}{eB/h} \right] \right) \frac{\hbar eB}{m} \quad (5.32)$$

So we can write with

$$x = \frac{\hbar eB/m}{E_f(0)} = \frac{eB/h}{N} \quad (5.33)$$

$$\frac{E_f(B)}{E_f(0)} = \left(\frac{1}{2} + \left[\frac{1}{x} \right] \right) x \quad (5.34)$$

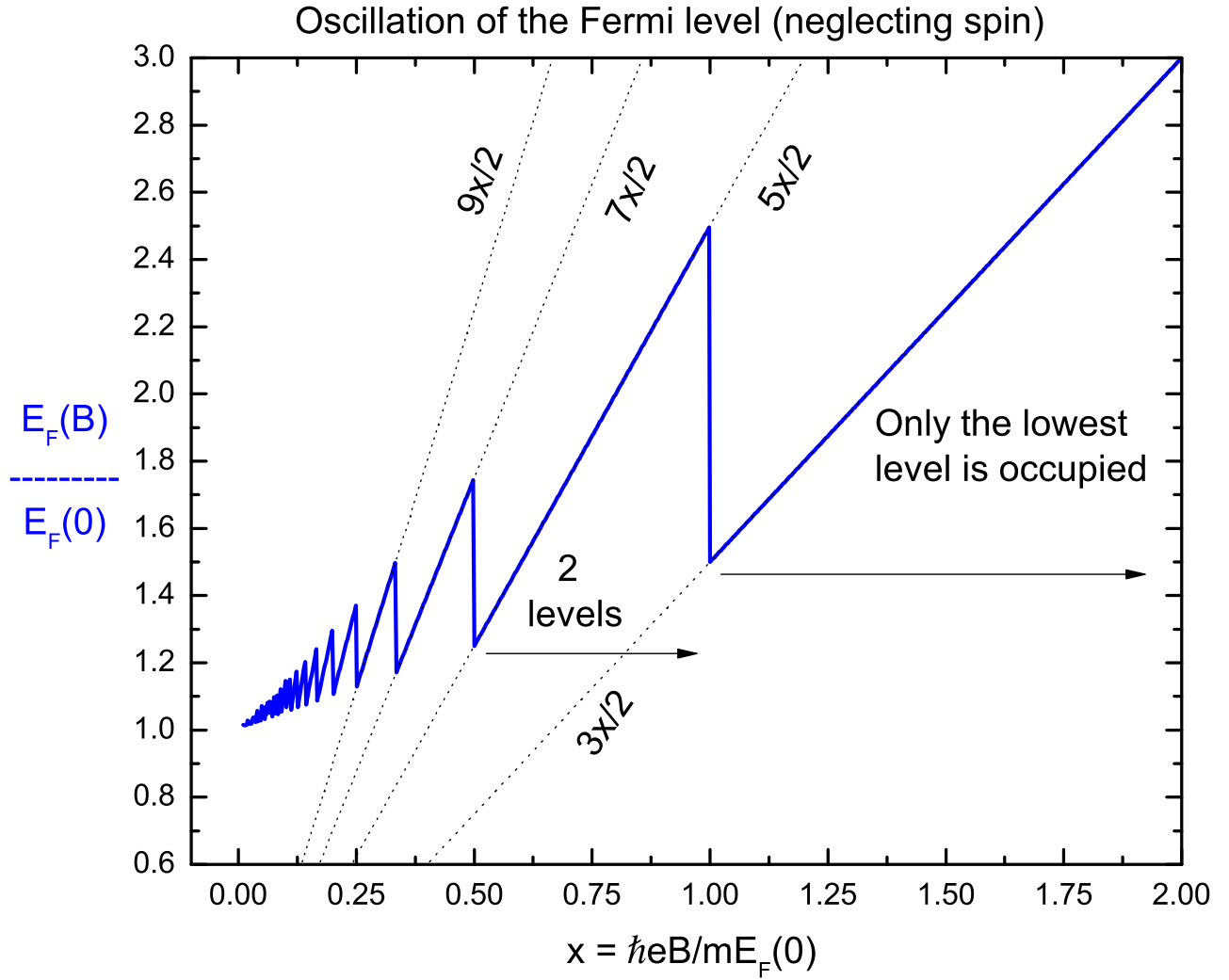


Figure 5.3: The Fermi level of the 2DEG (with its number density fixed) oscillates as the magnetic field is increased. Each level rises in energy and becomes more degenerate, linearly with increasing field. This kind of oscillation indeed occurs. It was experimentally demonstrated in a Si MOSFET by Pudalov et al *Sov. Phys. JETP* **62**, 1079 (1985)

5.2.3 The states near the edge of the sample

While calculating the degeneracy of each Landau level we didn't pay any particular attention to the small number of states that lie very close (within $\sim \ell$) to the edge. That doesn't introduce any error in the calculation because the number of such states can only scale as the perimeter whereas the number of states in the bulk (*i.e* sufficiently inside) of the $L \times W$ sample scales as the area.

But there is a time when the presence of such small number of states become important. That is the time when the Fermi level is moving from one Landau level to the next. Let us consider a situation where the field is just high enough to put all the particles in the lowest Landau level. Now, if we reduce the field a little then the degeneracy of the lowest level is no longer enough, so the Fermi energy (the highest occupied state at $T=0$) is expected to jump to the next higher state.

But now, recall that in any finite sample the edges must be like a potential barrier (see Fig.). A potential barrier appropriate for a sample with 2 edges, can be modelled by an "U-shaped" potential, with a flat bottom (the bulk region) and steeply rising edges containing the electrons from "falling out" of the sample. The functional form of the rise near the edges is unimportant. It is not necessary at all for the rise to be symmetrical on the two edges. As we have done earlier, we will assume that

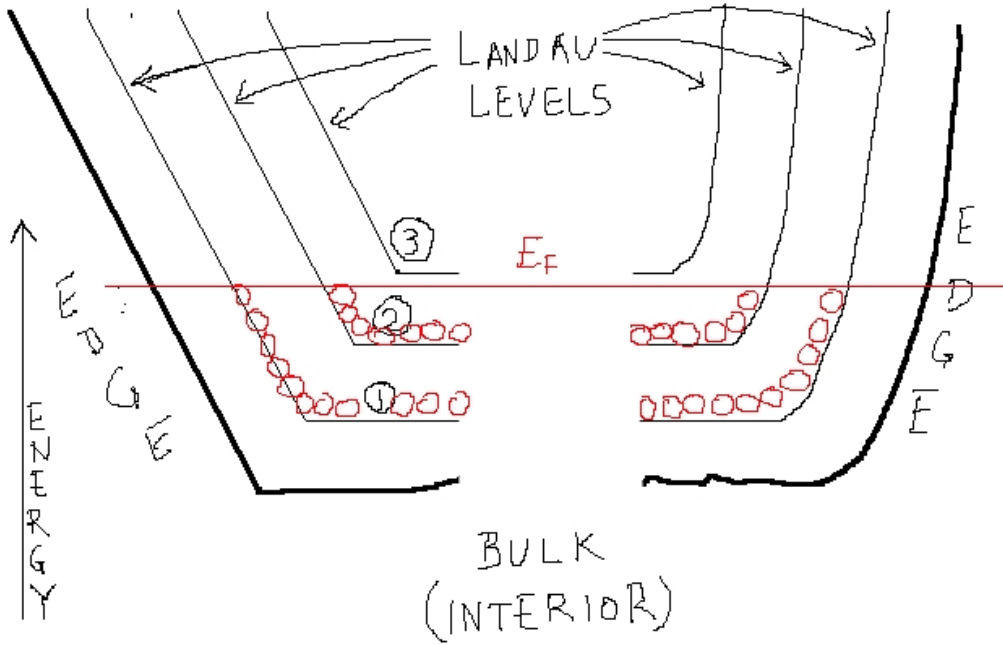


Figure 5.4: The energy of the states (Landau level) must rise as the location ($x_0 = k\ell^2$) approach the edges. The figure has been highly exaggerated to show this effect near the edges. But the important point is that when the Fermi energy tries to jump from one Landau level to another, it has to pass through a region when the density of states is zero in the bulk and the only current carrying states are in two narrow strips near to edges. These strips behave like one dimensional channels.

we can now make the eigenvalues position dependent. (see Fig.). But we exploit the fact that the position dependence of the (harmonic oscillator) eigenfunctions is related to the k_y .

$$E_n(x) = (n + 1/2)\hbar\omega + U(x_0) \quad (5.35)$$

$$\text{or } E_n(k) = (n + 1/2)\hbar\omega + U(k\ell^2) \quad (5.36)$$

So between jumping from one landau level to the next one, the Fermi energy moves through the states near the edge. In the bulk during this time the conductivity drop to zero.

Group velocity of the edge states

The group velocity is given by $v_g = \frac{1}{\hbar} \frac{dE}{dk}$. Notice that this implies v_g is zero the bulk. But as the energy of the states (due to the rise in the confining potential) rises near the edges, the derivative w.r.t k sees this rise, because $k = x_0/\ell^2$. This means that v_g must have opposite signs in the two edges. These are the left going and right going edge states.

Since there are no current carrying states in the bulk, and there is a strong magnetic field, both σ_{xx} and ρ_{xx} must now be very small, thus making the Hall angle δ very close to $\pi/2$. This is the quantum Hall state.

The narrow strips of current carrying regions are all that connects the ohmic contacts. There is a remarkable analogy between these states and one dimensional conduction that we will now see.

5.2.4 The "perfect" one-dimensionality of the edge states

The current crossing any transverse plane placed in the path of a 1D-channel can be written as the difference between the number of particles crossing from left to right ($I_{L \rightarrow R}$) and from right to left ($I_{R \rightarrow L}$).

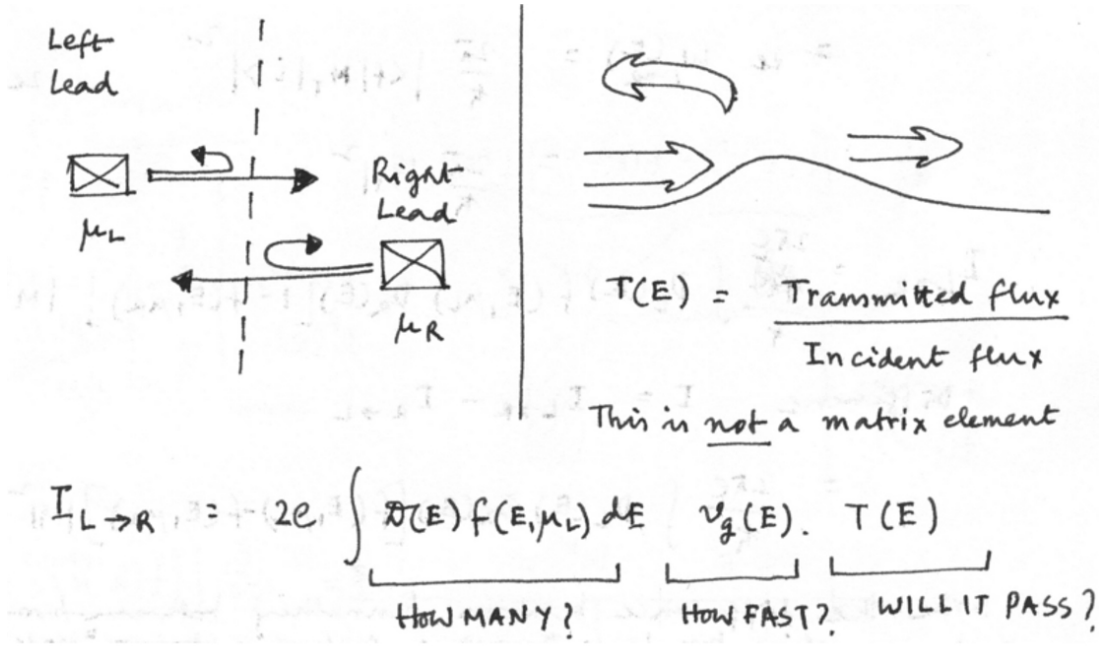


Figure 5.5: Explanation of the various terms in the expression for current. Note the similarity with the simple expression for current $j = env$.

$$\begin{aligned}
 I_{L \rightarrow R} &= e \int_{-\infty}^{\infty} dE D(E) f(E, \mu_L) v_g(E) T(E) \\
 I_{R \rightarrow L} &= e \int_{-\infty}^{\infty} dE D(E) f(E, \mu_R) v_g(E) T(E) \\
 I &= e \int_{-\infty}^{\infty} dE [f(E, \mu_L) - f(E, \mu_R)] v_g(E) T(E)
 \end{aligned} \tag{5.37}$$

- Some particles are launched from the left lead (L). They share the electrochemical potential of the left lead.
- We can write the current due to these as shown in the set of equations 5.37.
- Note that we have ignored the question whether there are empty states for these to go to in the other lead. The crucial assumption is that the particles will find a way (equilibrate) with the lead when they reach it. The leads are connected to external voltage source which will force its electrochemical potential.
- There is a subtle difference between the formula we just wrote and a similar looking formula for tunneling conductance that involves thin barriers etc.
- Also note that the factor $T(E)$ is the ratio of two fluxes, it is not a matrix element.

The difference between the electrochemical potentials can be written by taking $\mu_L = \mu - eV/2$ and $\mu_R = \mu + eV/2$. If this is small then we can write

$$f(E, \mu + eV/2) - f(E, \mu - eV/2) \approx \frac{\partial f(E, \mu)}{\partial \mu} eV = -eV \frac{\partial f(E, \mu)}{\partial E} \tag{5.38}$$

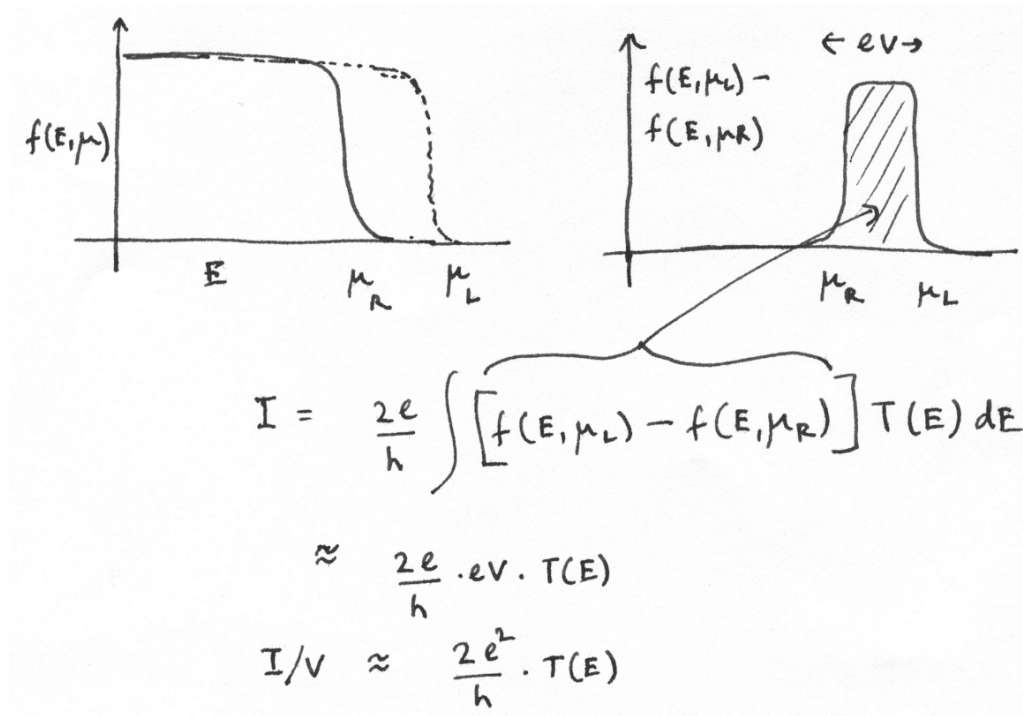


Figure 5.6: Which region contributes to the expression for the current integral?

because we know that in the Fermi function only the combination $E - \mu$ occurs. Note that the derivative of the Fermi function w.r.t. energy has again appeared quite naturally.

At $T = 0$ the difference between the Fermi functions on the two sides is non-zero only in the small interval between μ_L and μ_R . So we can simplify the eqn 5.37 to:

$$I = e \int_{\mu_R}^{\mu_L} dE D(E) v_g(E) T(E) \quad (5.39)$$

$$= e \int_{\mu_R}^{\mu_L} dE \left(\frac{1}{\pi} \frac{dk}{dE} \right) \left(\frac{1}{\hbar} \frac{dE}{dk} \right) T(E) \quad (5.40)$$

$$= \frac{2e^2}{h} (V_L - V_R) T(E) \quad (5.41)$$

Note how the fortuitous cancellation of the density of states factor and group velocity works - but *only* in 1 dimension, making it somewhat special. If we wrote an expression for conductance ($G = I/V$) in 2-dimensions, then the carrier concentration would have appeared in the expression, like $\sigma = ne\mu$, that we have seen many times. But the cancellation of the density of states with group velocity took care of that.

Now consider the factor $T(E)$. A left mover can be sometimes scattered to a right moving state in the channel, because they can physically occupy the same space. This means that $T(E)$ cannot be guaranteed to be 1, because there can be some backscattering.

In the hall bar, however, the edge channels have been pulled apart to the opposite edges of the hall bar. A edge state on the top (Fig. 5.7) has no amplitude to exist or to get scattered to the bottom edge. The left and right going states are separated by typically 10-100 microns.

Referring to Fig. 5.4, there can be more than 1 edge states on the same edge and some electrons may be scattered from one channel to the other. But this causes no loss of current as the particles keep

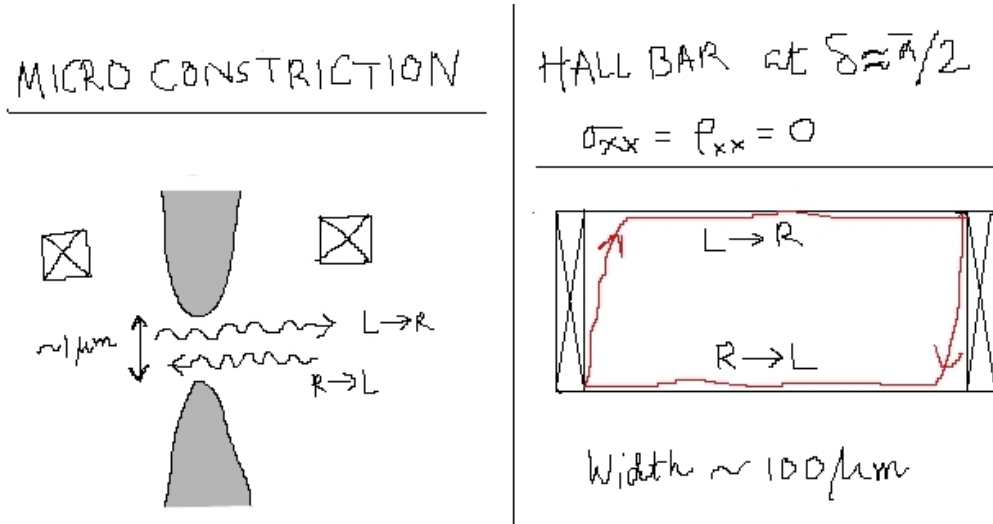


Figure 5.7: The essential difference between the 1D channel in a micro-constriction and the "edge states" of a Hall bar is that the left and right movers are spatially separated in the Hall bar by the width of the Hall bar - which may be even millimeters wide.

moving in the same direction. The contribution of a channel to the conductivity didn't depend on how many carriers are there in the channel.

Now, say we raise the potential of the left ohmic contact by an amount eV . The ohmic contact and the edge state emanating from it share the same chemical potential. In the quantum hall state, the potential difference between the current injecting contacts is same as the potential difference between the opposite edges - which is the Hall voltage. If the number of excess carriers injected is δn ,

$$\delta n = ND(E)e\delta V \quad (5.42)$$

$$I = (ev_g) [ND(E)e\delta V] \quad (5.43)$$

$$= \frac{e}{\hbar} \frac{dE}{dk} N \frac{1}{2\pi} \frac{dk}{dE} e\delta V \quad (5.44)$$

$$\therefore R_H = \frac{\delta V}{I} = \frac{h}{Ne^2} \quad (5.45)$$

An exact result independent of disorder and sample mobility.

5.2.5 Disorder broadening of the Landau levels

Over what extent in magnetic field would this special state exist? This is important because in reality that would determine how easy it is to observe the state. We have already calculated the degeneracy of each Landau level - but in reality these (eB/h) states are not concentrated in a zero width delta function. There are small random variations of the potential (bottom of the conduction band) throughout the sample that result in a small spread of the energy of the eigenstates. See Fig. 5.8 for the density of states would look in a real sample. For example in a GaAs-AlGaAs heterostructure the Landau level gap is about $20k_B/\text{Tesla}$ and the width of each state might be about $\sim 1k_B$, resulting from random disorder.

5.2.6 Longitudinal resistance and "Quantum" Hall effect

The exactness of the quantum Hall effect was discovered in 1980 [K. von Klitzing, G. Dorda and M. Pepper, *Physical Review Letters*, **45**, 494 (1980)] in Silicon MOSFETs. However the phenomena can be observed in any 2 dimensional electronic system (satisfying certain criteria). The data shown in

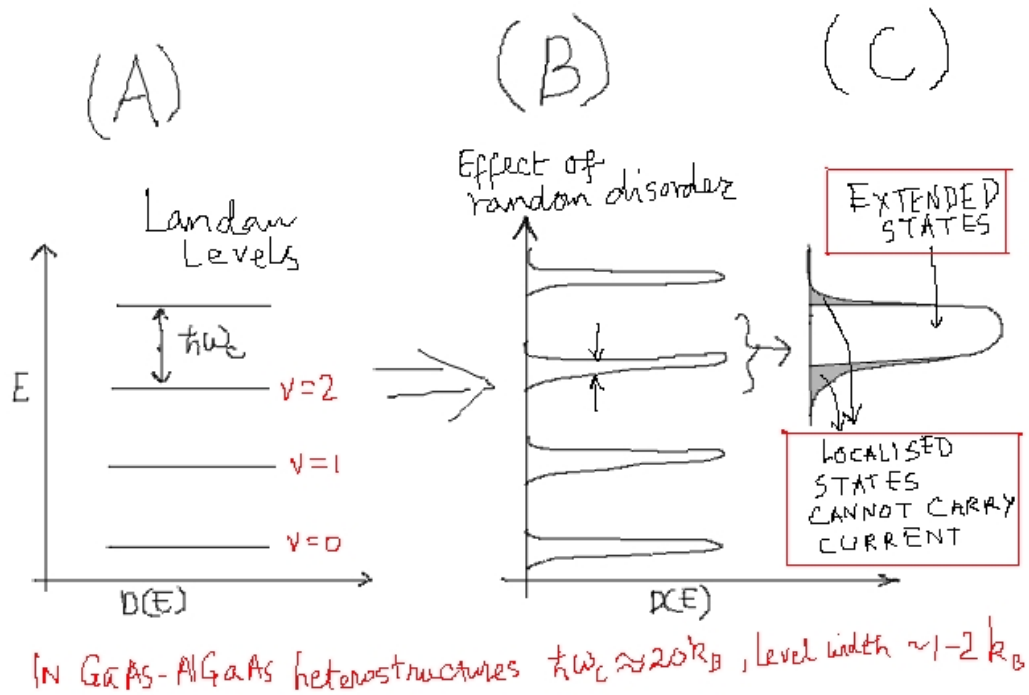


Figure 5.8: The effect of disorder makes each delta-function in the density of states broaden over an energy range equivalent to 1-2 Kelvin. It also turns out that the states which suffer maximum displacement and are pushed towards the band tails, are localised. They cannot carry current and when the Fermi level resides in these states the "bulk" continues to be non-conducting. When the Fermi level lies in the extended (free electron like) states, then the bulk of the sample conducts and the Hall voltage is not quantised - because the transport is not edge-state controlled.

Fig. 5.9 comes from a GaAs-AlGaAs heterostructure. The oscillations in the longitudinal resistance are called "Shubnikov de-Haas" oscillations and contain a wealth of interesting physics that you can read about in the references given at the beginning.

The information about how disordered the sample is hasn't been completely lost though. You can reason it out that the width of the Hall plateau (the interval in magnetic field over which the plateau exists) is a measure of the disorder. Large width means lower mobility! In fact in extremely "clean" samples the width of the Hall plateaus will become *very small* - a somewhat counter-intuitive result.

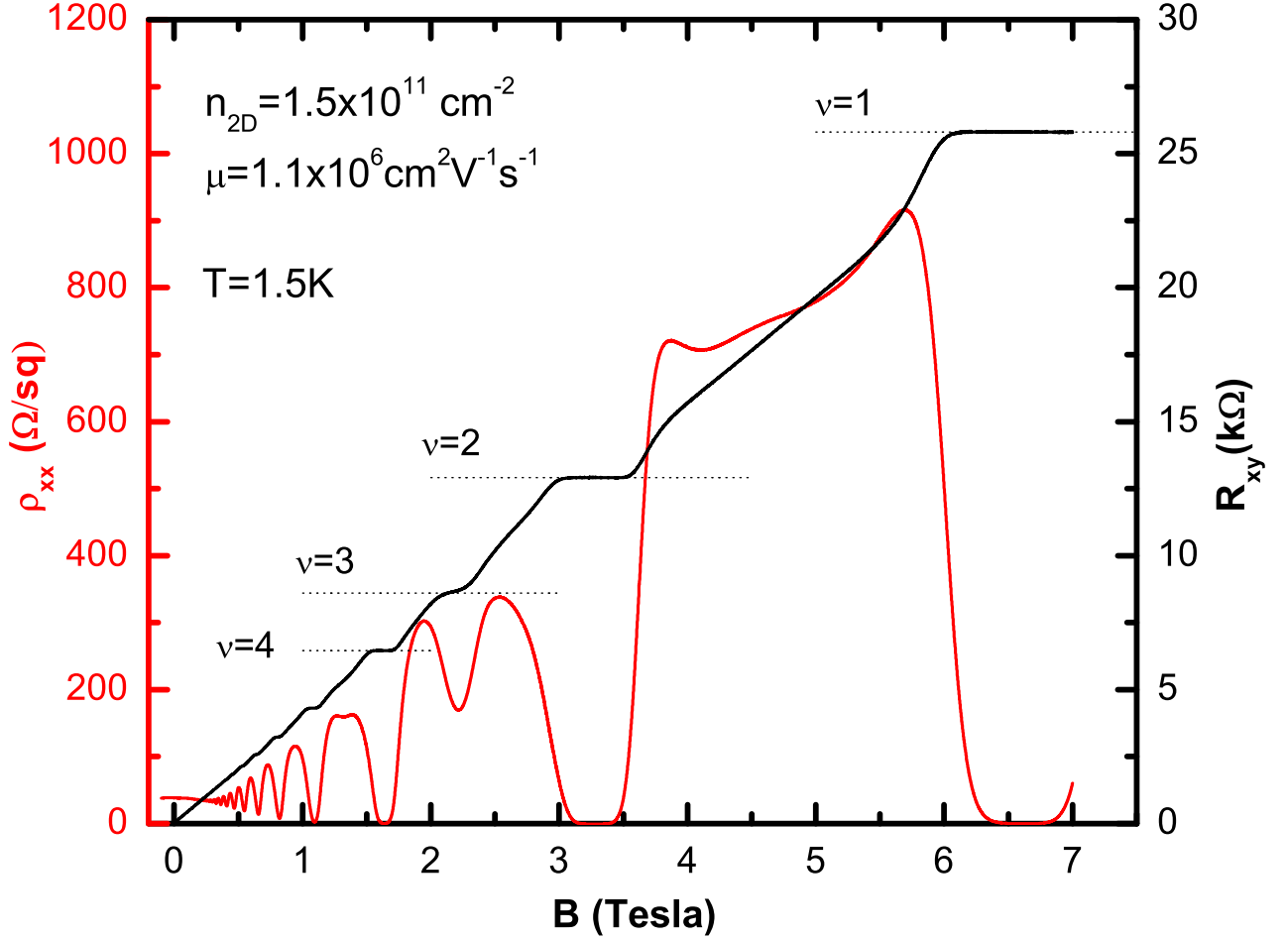


Figure 5.9: Experimental data from a GaAs-AlGaAs heterostructure, showing the longitudinal resistance (red) and hall resistance (black). Notice that the zeros of ρ_{xx} coincide with the quantum hall plateaus. These are the times when the Fermi level lies in the localised states in the band tails. The fact that ρ_{xx} and σ_{xx} are simultaneously zero comes from the nature of the conductivity matrix, as we have discussed before. The numbers along with the plateaus denote how many Landau levels are filled at that point, and hence the number of edge states in the system as well.

Chapter 6

Conduction through a constriction: quantisation of conductance

References:

1. D.A. Wharam, T.J. Thornton, R.. Newbury, M. Pepper, H. Ahmed, J.E.F. Frost , D. G. Hasko, D.C. Peacock, D.A. Ritchie AND G.A.C. Jones, *J. Phys. C*, **21**, L209 (1988)
2. B.J. van Wees, J. van Houten, C.W.J. Beenaker, J.G. Williamson, L.P. Kouwenhoven, D. van der Marel and C.T. Foxon, *Phys. Rev. Lett.*, **60**, 848 (1988)

We saw in the last set of lectures that the remote doping method can lead to very high mobilities. This means that due to long mean free paths, we can think of interference between electron paths and ballistic conduction. In this lecture we will see how (using a GaAs-AlGaAs heterostructure) one can design a narrow constriction through which electrons may be transmitted without scattering. This is possible because in GaAs-AlGaAs heterostructures the mean free path of the electrons can easily be about ~ 10 microns, enabling lithographic gates and other features to be of smaller size than the mean free path. A striking "quantum" consequence of this is that at low temperature the quantisation of conductance of a constriction. Typically such a constriction would be ~ 0.5 micron in length and of the order of Fermi wavelength in width. This effect was first experimentally shown in 1988-89.

6.1 How is such a constriction made?

Microconstrictions are now-a-days easily made by photolithography and electron-beam lithography. The following pictures should give you an idea about how these are made (explanations of the steps in lecture).

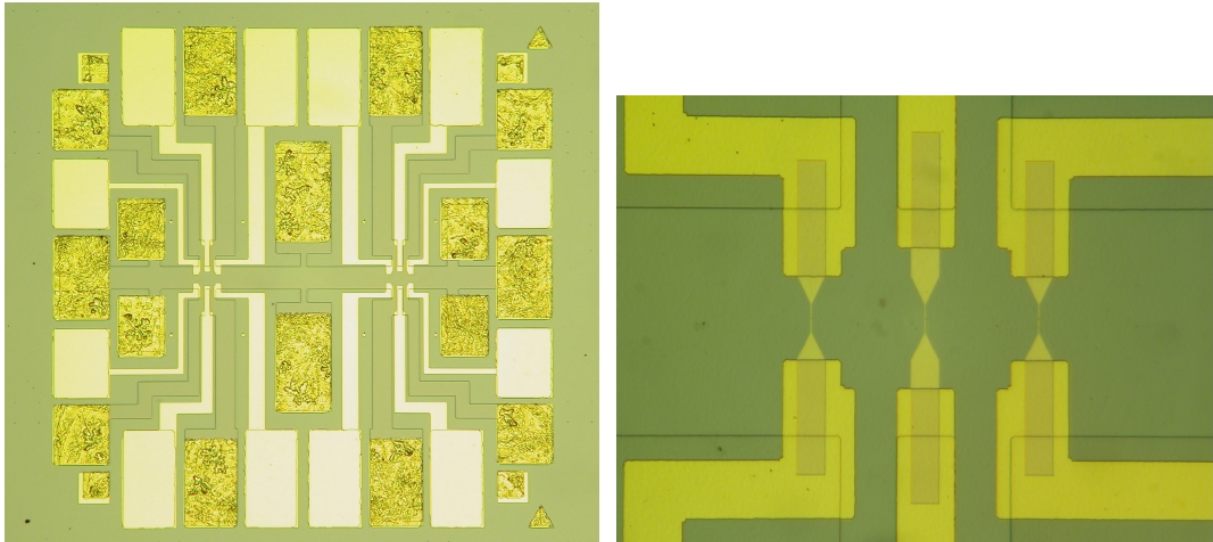


Figure 6.1: (left) Hall bar with optical gates which connects to the finer features later on. The hall bar is about 1mm long. (right) Expanded view of the split-gate region

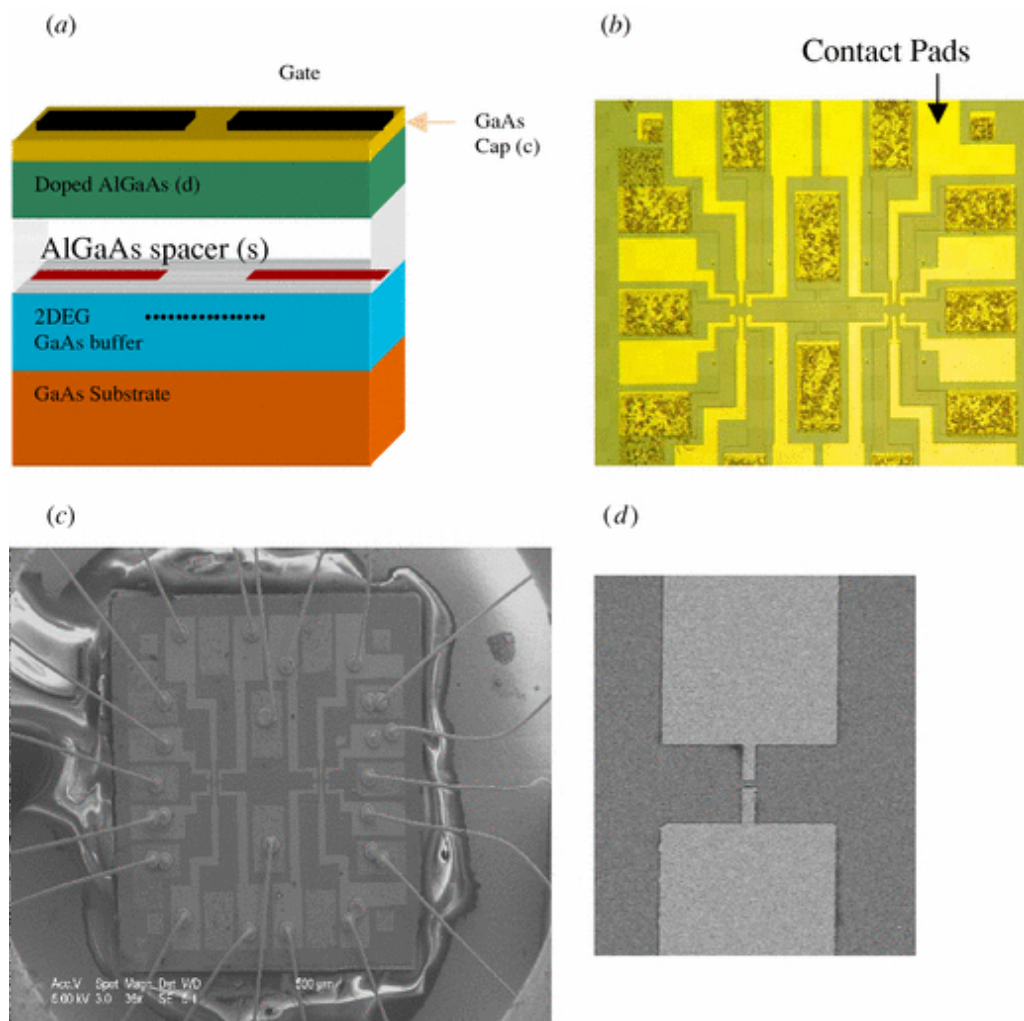


Figure 6.2: Another picture describing the process.

6.2 Calculation of the conductance of a Quantum point contact (QPC)

Assume that the electrochemical potential drop is only over the constriction. The 2DEG on both sides are unperturbed and their populations follow the Fermi distribution. The electrochemical potentials differ by an amount eV across the QPC. The current density \mathbf{j} can be written in general as

$$\mathbf{j} = ne\langle \mathbf{v} \rangle \quad (6.1)$$

However here \mathbf{v} is not the *drift* velocity. Because we are dealing with ballistic conductance, we can take the velocity components from the Fermi circle. This is an important point to notice.

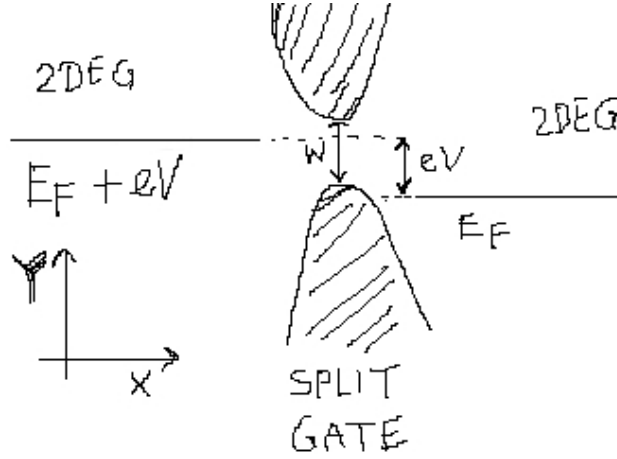


Figure 6.3: Schematic of the split gate, the effective width of the constriction will depend on the gate voltage.

The number of state contributing to this conduction process will be

$$n = D(E_F)eV = \frac{m}{\pi\hbar^2}eV \quad (6.2)$$

Hence we can write for the conductance (G) using eqn 6.1 (we need to retain only j_x) as

$$j_x = \frac{m}{\pi\hbar^2}eV \cdot e \cdot \frac{\hbar}{m} \langle k_x \rangle \quad (6.3)$$

$$\therefore G = \frac{W j_x}{V} \quad (6.4)$$

$$= \frac{e^2}{\pi\hbar} W \langle k_x \rangle \quad (6.5)$$

We will average over the full Fermi circle with a factor of 1/2 to take care of the fact that only half of the electrons there contribute to current in the $+x$ direction.

Now, we make the crucial assumption that only those wavevectors are allowed to pass for which k_y is such that

$$k_y = \frac{n\pi}{W} \quad \text{where} \quad (6.6)$$

$$n = \pm 1, \pm 2, \dots \quad (6.7)$$

We assume that these modes will not mix (scatter from one to the other) as they pass through the constriction.

The Fermi surface average $\langle k_x \rangle$ is :

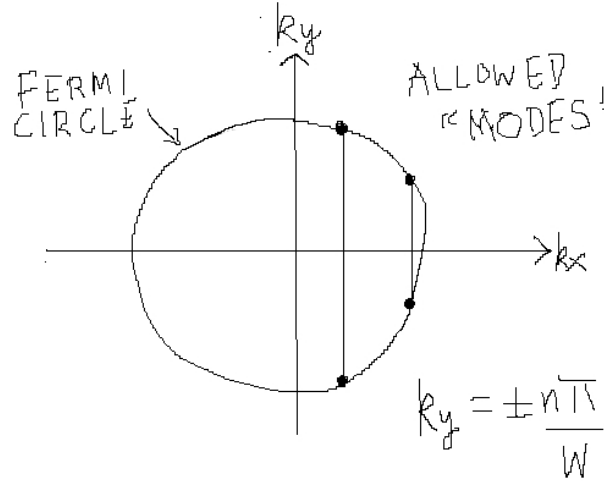


Figure 6.4: The allowed modes come from the right going part of the Fermi circle and the quantisation condition is set by $k_y = \pm \frac{n\pi}{W}$. For a typical density of $n \approx 1 \times 10^{11} \text{ cm}^{-2}$, the Fermi wavelength $\lambda_f = 2\pi/k_f$ would be about 100nm.

$$\langle k_x \rangle = \frac{1}{2\pi} \int_0^{2\pi} d\theta |k_f \cos \theta| \sum_{-\infty}^{\infty} \delta \left(\frac{W k_f \sin \theta}{\pi} - n \right) \quad (6.8)$$

$$= \frac{1}{2W} \int_0^{2\pi} d\theta \frac{W k_f \cos \theta}{\pi} \sum_{-\infty}^{\infty} \delta \left(\frac{W k_f \sin \theta}{\pi} - n \right) \quad (6.9)$$

$$= \frac{2}{W} \int_0^{\pi/2} d\theta \frac{W k_f \cos \theta}{\pi} \delta \left(\frac{W k_f \sin \theta}{\pi} - n \right) \quad (6.10)$$

$$= \frac{2}{W} \int_0^{W k_f / \pi} dx \delta(x - n) \quad (6.11)$$

$$= \frac{2}{W} \left[\frac{W k_f}{\pi} \right] \quad (6.12)$$

Now remembering the factor of 1/2 and using eqn 6.12, where $[]$ denotes the largest integer contained in the expression, eqn 6.5 gives

$$G = \frac{2e^2}{h} \left[\frac{W k_f}{\pi} \right] \quad (6.13)$$

Note the following points:

- Thus G is quantised in units of $\frac{2e^2}{h} = 78\mu S$, which is approximately $12.8k\Omega$ in resistance units.
- If W , the width of the constriction, is very small ($W < \pi/k_f$), then no channel can come into conduction. Only when W and the Fermi wavelength (λ_f) start becoming comparable, conduction starts.
- If W is very large, then we get the classical limit (called the Sharvin formula),

$$G_{\text{classical}} = \frac{2e^2}{h} \frac{W k_f}{\pi} \quad (6.14)$$

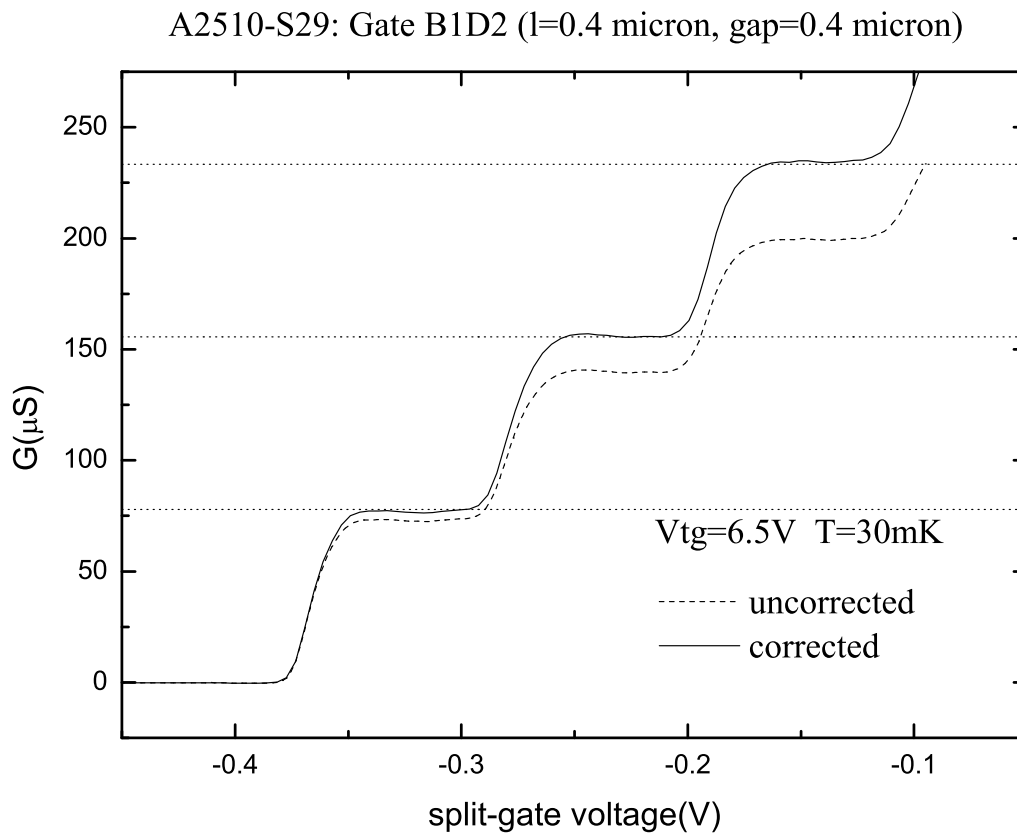


Figure 6.5: The measured conductance of a split gate at low temperatures ($T=30\text{mK}$). Note the quantisation of conductance. The sharpness of the quantisation will be smeared out at higher temperatures.

1-d channel in a magnetic field

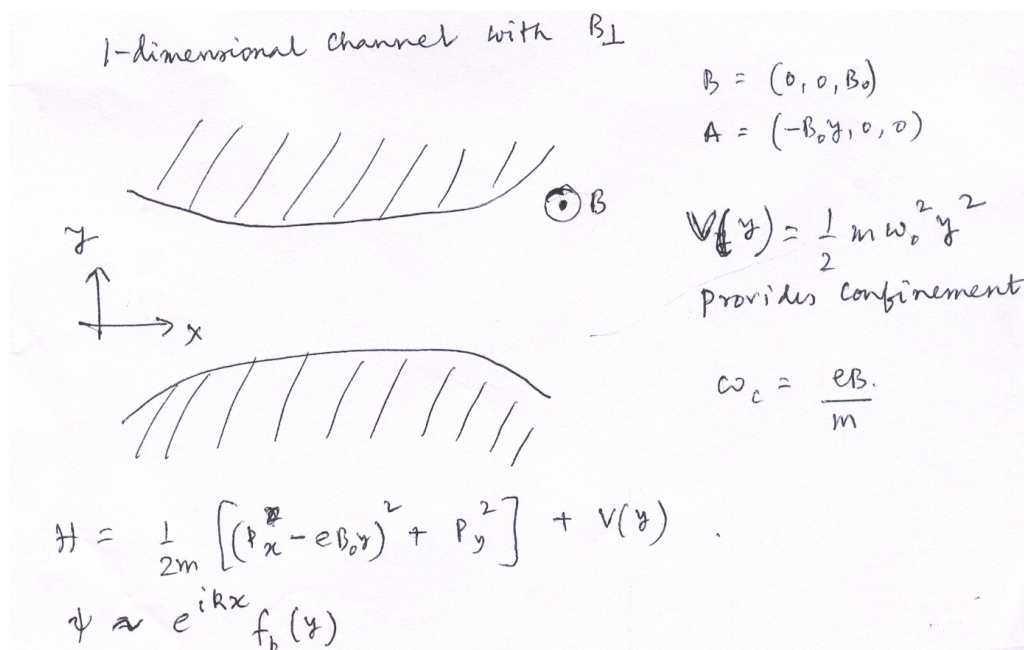


Figure 6.6: The orientation of the channel and the magnetic field.

PROBLEM : The energy of an electron in a 1-d channel is the sum total of the energy of confinement in one direction (like particle in a box) and the kinetic energy of travelling wave in the direction of motion. So the allowed energies are

$$E_n = \frac{\hbar^2}{2m} \left[k_x^2 + \left(\frac{n\pi}{W} \right)^2 \right] \quad (6.15)$$

if the confinement is taken to be a hard wall type. Here W is the width of the channel and for each value of n we have $n - 1$ nodes and these are called successive subbands. However for the rest of the problem, consider the confinement to be parabolic (harmonic oscillator type) in the y direction. This is easier to handle mathematically. You should be able to write down the allowed energies at once. Next, consider a 1-d channel placed in a perpendicular magnetic field as shown. The confinement is provided by a parabolic potential (electrostatic). The confining potential, the magnetic vector potential and the wavefunction can be written as

$$V(y) = \frac{1}{2} m \omega_0^2 y^2 \quad (6.16)$$

$$\mathbf{A} = (-By, 0, 0) \quad (6.17)$$

$$\psi(x, y) \sim e^{ikx} f_k(y) \quad (6.18)$$

Show that under these conditions the energy levels are given by:

$$E_n = \frac{\hbar^2 k^2}{2m} \left(\frac{1}{1 + (\omega_c/\omega_0)^2} \right) + \left(n + \frac{1}{2} \right) \hbar \sqrt{\omega_c^2 + \omega_0^2} \quad (6.19)$$

where $\omega_c = eB/m$ is the cyclotron frequency. Notice that near $k = 0$ the magnetic field raises the energy of the subband bottom, by making the apparent confinement steeper.

Try to solve it for a magnetic field that is in plane and parallel to the channel.

Try to solve it for a magnetic field that is in plane and perpendicular to the channel.

Chapter 7

Superconductivity

References:

1. *Superconductivity of metals and alloys*, P. G. de-Gennes. W. A Benjamin publishers (1966)
 2. *Statistical mechanics* R. P. Feynman. Advanced Book classics. (2nd ed, 1998). See also the *Feynman lectures, vol 3* for a special lecture on superconductivity.
 3. *Introduction to superconductivity*, M. Tinkham, McGraw Hill publishers. This is a standard and extensive textbook for superconductivity.
 4. *Solid State Physics*, chap 18, by G. Grosso and G.P. Parravicini, Elsevier (Singapore).
-

We will try to present a very brief account leading to the microscopic theory. Ultimately we want to be able to see what happens at the junction between two superconductors, in a not too handwaving manner. We will show most of the important steps of the argument, but not try to deduce everything.

The experimental discovery of superconductivity (zero resistance state in Mercury) and Bohr's analysis of the hydrogen atom happened around the same time. However nearly 50 years of development of quantum mechanics and statistical physics was needed before the explanation of this striking phenomena emerged. But superconductivity is not a very rare phenomena, infact there are more naturally superconducting elements (in the periodic table) than ferromagnetic elements.

7.1 What do we know experimentally

7.1.1 It is a distinct thermodynamic phase

1. Some materials (like Hg, Sn, Al, Nb and many more) loose their electrical resistance when cooled below a certain temperature. Among naturally occurring elements (under atmospheric pressure) Nb has the highest T_C , 9.3K. There are quite a few in the range 1-10K.
2. Interestingly very good conductors like Au, Cu, Ag do not become superconducting.
3. Superconductivity is distinct thermodynamically from infinite conductivity. The following experiment tells us why:

Consider a very long coil of radius R and N turns/mt, surrounding a very long cylinder of a superconducting material of radius r . The setup is initially above T_C . A current I is maintained in the coil by a constant current source.

- Initially there is an uniform axial field $B = \mu_0 NI$ in the coil and we assume that it is the same inside the material also. It amounts to saying that the magnetic permeability of the material is 1, which is often a very good approximation.
- We find that below T_C the superconductor expels the magnetic field, provided the magnetic field is less than a critical value $B < B_C$.
- This means that the flux through the coil has changed by

$$\Delta\phi = \pi r^2 B_C N \quad (7.1)$$

- Let us consider the Helmholtz free energy of the system before and after the transition. During the time the flux was changing, because the current was kept constant, the system did some work on the external (battery) that was maintaining the current. We have by adding the energy of the magnetic field and the material, (all quantities are to be interpreted as per unit length)

$$F_i = F_n + \pi R^2 \frac{B_C^2}{2\mu_0} \quad (7.2)$$

$$F_f = F_s + \pi(R^2 - r^2) \frac{B_C^2}{2\mu_0} \quad (7.3)$$

$$\int_{F_i}^{F_f} dF = I \int_{\phi_i}^{\phi_f} \frac{d\phi}{dt} dt \quad (7.4)$$

$$F_f - F_i = NI(-\pi r^2 B_C) \quad (7.5)$$

$$= -\pi r^2 B_C \left(\frac{B_C}{\mu_0} \right) \quad (7.6)$$

$$\therefore F_s - F_n = -\frac{\mu_0 H_C^2}{2} \quad \text{in SI} \quad (7.7)$$

$$= -\frac{H_C^2}{8\pi} \quad \text{in cgs} \quad (7.8)$$

- The free energy difference between the two phases is thus reduced to an easily measurable quantity, the critical field. The various derivatives of free energy can now be used to compute many relevant quantities (entropy, latent heat and specific heat..)

$$\Delta F = F_n - F_s = -\frac{\mu_0 H_C^2}{2} \quad (7.9)$$

$$\therefore S_n - S_s = -\frac{d\Delta F}{dT} = -\mu_0 H_C \frac{dH_C}{dT} \quad (7.10)$$

$$\Delta L = T(S_n - S_s) = -\mu_0 H_C T \frac{dH_C}{dT} \quad (7.11)$$

$$\Delta C_v = C_n - C_s = -\mu_0 T \left(\frac{dH_C}{dT} \right)^2 \quad (7.12)$$

PROBLEM : Complete the derivations of the equations given above

- Note that the superconducting transition is of second order (no latent heat) in zero magnetic field.
- Experiments show that the quantity $F_n - F_s$ is of the order of $\frac{(k_B T)^2}{E_F}$ per electron, which is clearly very small compared to E_F since the temperatures involved are a few kelvin only. Thus the superconducting state differs from the normal state by a very small amount of energy.

- The energy involved must be of the order of $k_B T_c$. At 10K, this is less than 1 meV. In frequency units, this would mean about 10^{11} Hz.
- Superconductivity is not the same as perfect conductivity in the sense that we cannot get the flux expulsion result (Meissner effect) by putting $\sigma \rightarrow \infty$ in Maxwell's equations. That would tell us that the flux through a loop of infinite conductivity should stay frozen. *You should be able to prove this result.*

See Fig. 7.1 for the specific heat change at the transition. The electronic specific heat of the normal metal is linear with T as expected for free electrons obeying Fermi statistics.

PROBLEM : Why is an exponential specific evidence of a gap? Consider a two level system with two energy levels $E_1 = 0$ and $E_2 = E$. The partition function is simple to write it is $Z = 1 + e^{-\beta E}$. Now calculate the Helmholtz free energy, F , and then C_v from the second derivative of F , using standard thermodynamic relations. Remember C_v must be positive. You should get

$$\frac{C_v}{k} = \left(\frac{E}{kT} \right)^2 \frac{e^{-\beta E}}{(1 + e^{-\beta E})^2} \quad (7.13)$$

How does this behave at low temperatures? If there is a gap just above the ground state, then at low temperature we can get an idea of the system's behaviour by just considering the two lowest states as we have done here.

7.1.2 Evidence for a gap

- Absorption of radiation: Superconductors absorb microwave radiation of $\lambda \sim 1\text{mm}$. See Fig.
- Tunneling experiments show that the density of states has a gap.

See fig. 7.3 to get an idea what tunneling experiments tell us.

These two taken together suggest that pairing plays a role, because the gap that we get from tunneling experiments is half the gap shown by the radiation absorption experiments.

Significance of a gap

A gap in the energy spectrum means that low energy excitations cannot be absorbed by the system or the system cannot give up small amounts of energy. This needs to be understood very carefully. Every bandgap does not lead to superconductivity - if it did, then a full valence band (as in intrinsic semiconductors) would become superconducting with a huge (few eV) gap! Clearly that does not happen.

- We must be able to build a current carrying configuration out of states just below the gap. That isn't possible if we use all the electrons in a band. If we displace a full band of electrons then we don't get any net current, because the sum of the group velocities of all the states over one full period (Brillouin zone) is always zero.
- An overall shift of the occupied states in k-space is needed to build a current carrying state. This is possible to do in a metal or something that has a partially filled band. Once this has been done the presence of a gap prevents the current carrying state from decaying in small bits and pieces. Of course if sufficient energy (in the form of kinetic energy of the critical current, or radiation etc) is available then a decay of that state will occur.
- From experiments we find that the gap varies with temperature and at $T = 0$ has a value

$$2\Delta(T = 0) = 3.52 k_B T_C \quad (7.14)$$

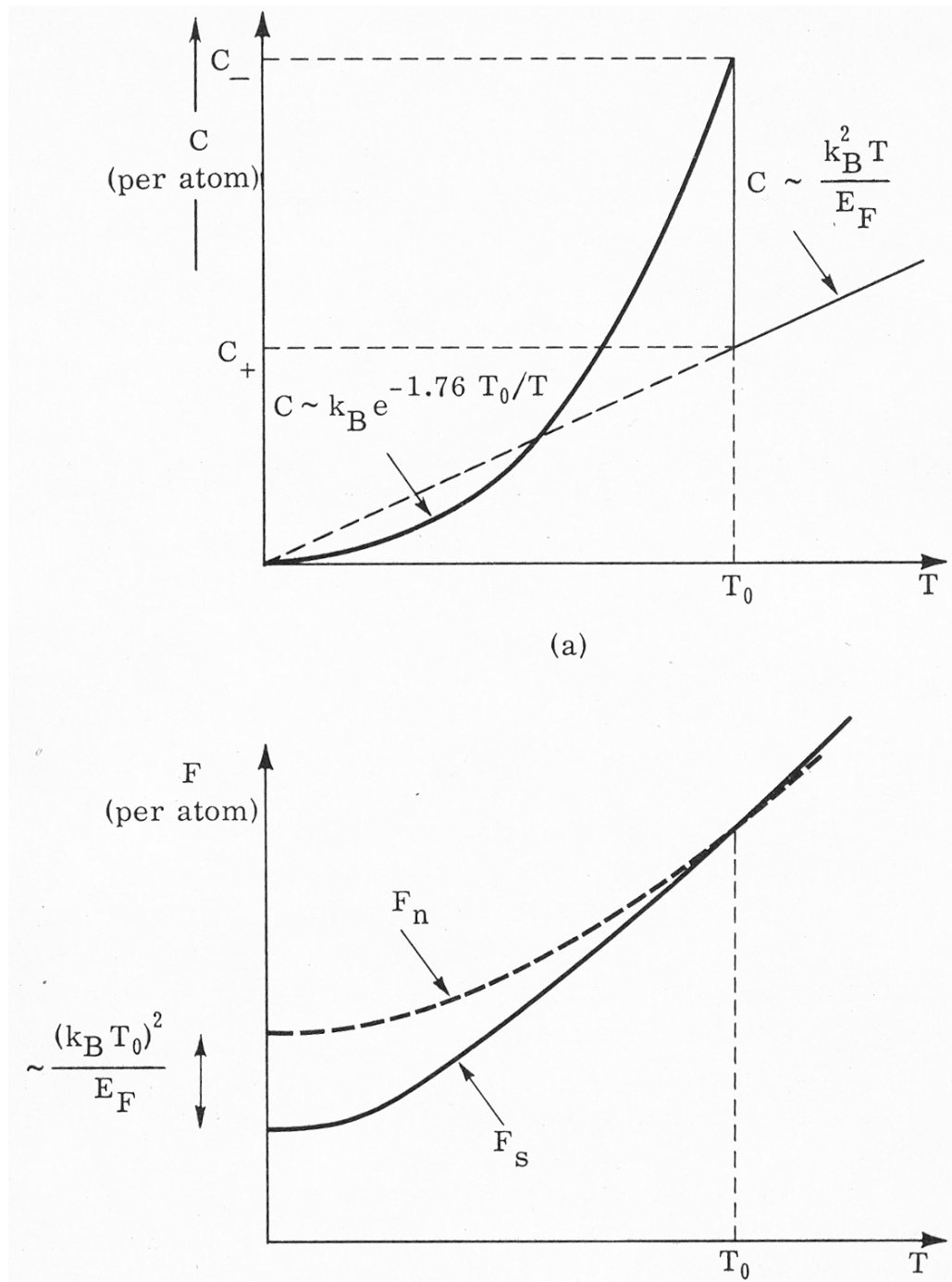


Figure 7.1: The change in specific heat as one passes the transition. The figure is taken from P.G. de-Gennes's book.

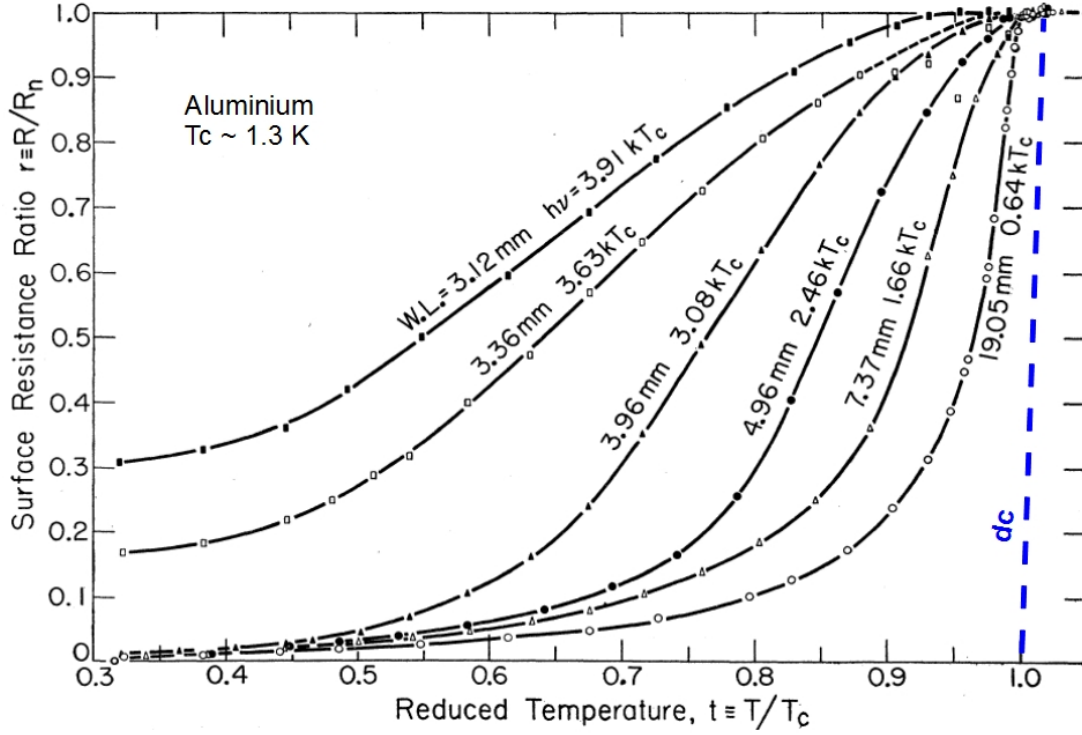


Figure 7.2: Millimeter wave absorption data from "Millimeter Wave Absorption in Superconducting Aluminum. I. Temperature Dependence of the Energy Gap", by M. A. Biondi and M. P. Garfunkel. *Physical Review B*, **116**, 853 (1959). The curve obtained at the longest wavelength, 19.05 mm, resembles the lower frequency microwave data in that there is a rather abrupt decrease in the absorption as the temperature falls slightly below T_c , and the absorption tends essentially to zero as t approaches zero. At successively shorter wavelengths (higher photon energies) the absorption falls more and more slowly as the temperature is reduced below T_c . However, for all photon energies less than $3.08k_B T_c$ the absorption electively approaches zero as $t \rightarrow 0$. In terms of an energy gap picture, the absence of absorption at absolute zero indicates that these photon energies ($h\nu < 3.08k_B T$) are too small to span the gap. Similarly, the substantial absorption at absolute zero indicated by the curves for $h\nu > 3.63k_B T_c$, indicates that these photon energies induce transitions of electrons across the forbidden gap. I have added the blue dotted line to show the expected behaviour of the resistance, if the measurement was done with zero frequency (dc).

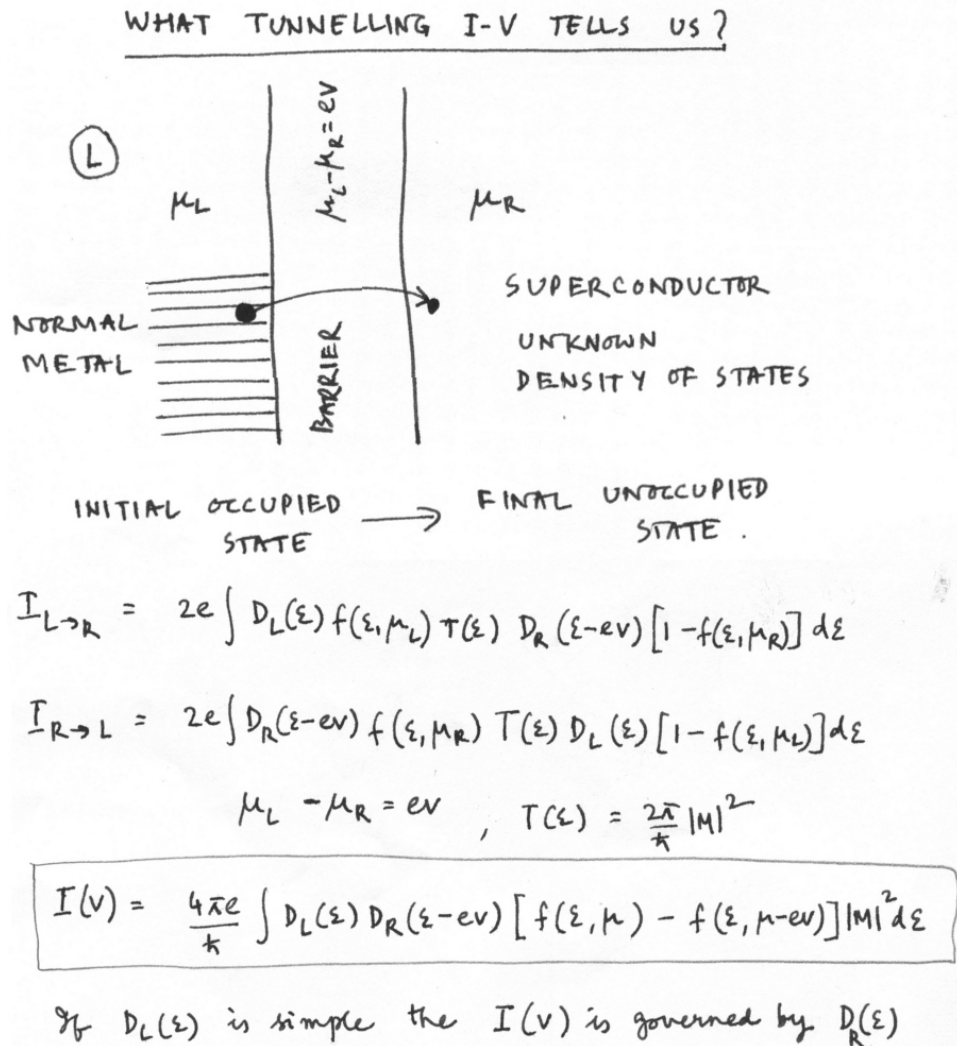


Figure 7.3: What does a tunneling experiment tell us? If the density of states on one side is simple then the resulting I-V is predominantly controlled by the density of states of the more complex side.

7.1.3 How an attractive e-e interaction becomes possible

We have encountered screening before. It tends to reduce the interaction strength between charges in a medium. Now could there be situations where the sign of the interaction gets changed? The presence of electrons in a polarisable lattice can give rise to such a possibility if we consider the screening of a time varying field. Technically it is called dynamic screening. Let us give a plausibility argument. *This is not a proof.*

Consider a system of metallic electrons in a lattice. We introduce a little extra charge $\delta\rho$ and as a result the charge density of the electrons change by an amount ρ_e and the charge carried by the ions (lattice) ρ_l . We must have

$$\nabla^2 V = -\frac{\delta\rho + \rho_e + \rho_l}{\epsilon_0} \quad (7.15)$$

Each ion of mass M carries a charge $(+Ze)$ and there are n per unit volume. So we can write the

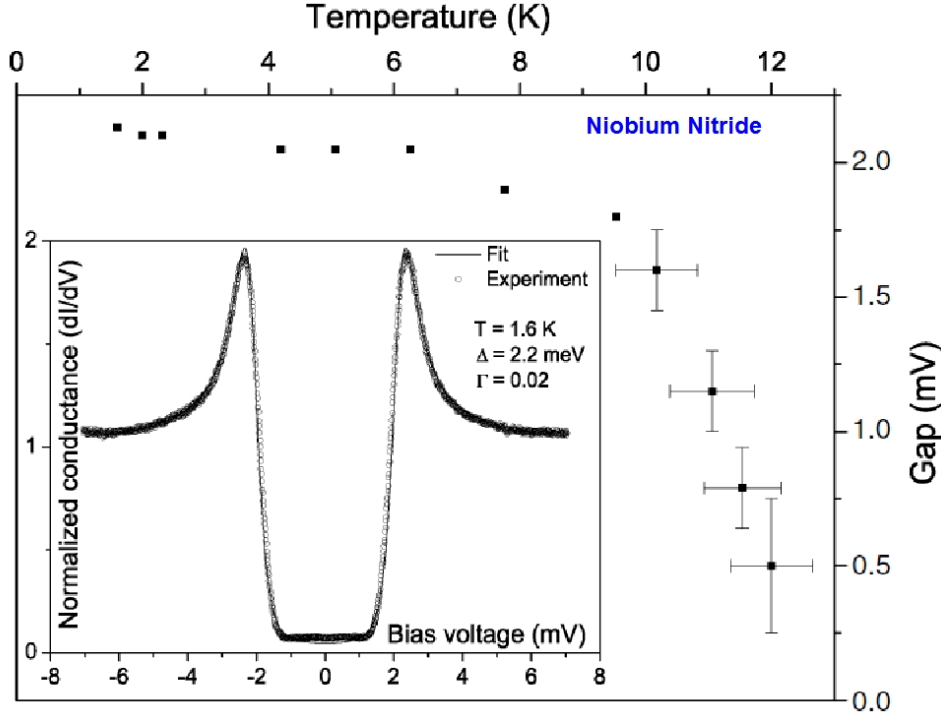


Figure 7.4: Observation by LT-STM of the superconducting energy gap variation with temperature in a 5-nm-thick NbN epi-layer deposited on R-plane sapphire. Inset: measurement at 1.6K of the quasiparticle density of state versus voltage (or energy) indicating completely empty electronic states below 1.5 meV and of the clear superconducting energy gap at 2.38 meV. The gap is well fitted by the BCS model with a pair-breaking parameter $\Gamma = 0.02$. The data is from "Fabrication of a superconducting niobium nitride hot electron bolometer for single-photon counting". by R Romestain, B Delaet, P Renaud-Goud, I Wang, C Jorel, J-C Villegier and J-Ph Poizat, in *New Journal of Physics* **6**, 129 (2004).

equation of motion of the ions as:

$$M \frac{d\mathbf{v}_l}{dt} = e\mathbf{E} \quad (7.16)$$

$$\therefore \frac{d\mathbf{j}_l}{dt} = \frac{nZe^2}{M} \mathbf{E} \quad (7.17)$$

But ion current and ion charge density must be related by the continuity equation. So

$$\begin{aligned} \frac{\partial \rho_l}{\partial t} &= -\nabla \cdot \mathbf{j} \\ \frac{\partial^2 \rho_l}{\partial t^2} &= -\frac{nZe^2}{M} \nabla \cdot \mathbf{E} \\ &= \frac{nZe^2}{M\epsilon_0} (\delta\rho + \rho_e + \rho_l) \end{aligned} \quad (7.18)$$

Notice that a plasma frequency of the lattice ions has emerged, call $\omega_l = \sqrt{\frac{nZe^2}{M\epsilon_0}}$. Now if we assume that the external potential has a time variation $\sim e^{i\omega t}$. Further we approximate the response of the electrons by its static response - like the Thomas Fermi case. The justification is that the lattice plasma frequency is much smaller than the electron plasma frequency. With these approximations we get

$$\omega^2 \rho_l(q, \omega) = \omega_l^2 [\delta\rho(q, \omega) + \rho_e(q, 0) + \rho_l(q, \omega)] \quad (7.19)$$

Using the Thomas-Fermi screening (k_{TF} is the Thomas-fermi wavevector) that we had introduced earlier we can write the relation between the electronic charge density and the total density as

$$\rho_e = \frac{k_{TF}^2}{q^2}(\rho_e + \rho_l + \delta\rho) \quad (7.20)$$

We now have enough information to find the ratio

$$\varepsilon(q, \omega) = \frac{\delta\rho}{\rho_e + \rho_l + \delta\rho} \quad (7.21)$$

It is left as an exercise to show that the result is

$$\frac{1}{\varepsilon(q, \omega)} = \frac{1}{1 + \frac{q_{TF}^2}{q^2}} \left(1 + \frac{\omega_q^2}{\omega^2 - \omega_q^2} \right) \quad (7.22)$$

where we have written

$$\omega_q^2 = \frac{q^2}{k_{TF}^2 + q^2} \omega_l^2 \quad (7.23)$$

Notice that there is the possibility of the screening changing sign for some range of frequencies. Thus the scattering between two electrons will *look like* as if they are via an attractive potential. If the phonon part is neglected then we get back the familiar Thomas-Fermi result.

Meissner effect and the gap

Let us for the moment assume that *somehow* the superconductor can be described by a wavefunction and the standard expressions for current apply. We will have more to say on this later. Following the argument of London we can write , (since there is no dissipation)

$$m\ddot{\mathbf{x}} = -e\mathbf{E} \quad (7.24)$$

$$\dot{\mathbf{j}} = -ne\dot{\mathbf{x}} \quad (7.25)$$

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} \quad (7.26)$$

The set of equations imply that

$$\frac{d\mathbf{j}}{dt} = -\frac{ne^2}{m} \frac{\partial \mathbf{A}}{\partial t} \quad (7.27)$$

we can then integrate this with the const of integration set to zero. We must then satisfy the boundary condition that $\mathbf{j}_\perp = 0$ at the boundary, so that $\mathbf{A}_\perp = 0$ as well. This choice of "guage" is called the transverse gauge. You can show that this (along with $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$) leads to

$$\nabla^2 \mathbf{A} = \frac{\mathbf{A}}{\lambda_L^2} \quad (7.28)$$

where $\lambda_L = m/ne^2$ has dimensions of length. The solution of the equation is clearly an exponential decay of the field as one enters the superconductor with a characteristic decay length (called the penetration depth) given by λ_L . This is a number of the order of few hundred angstroms for superconducting metals.

7.2 Ginzburg Landau "order parameter" theory

Microscopic theories start by trying to write down the wavefunction containing the co-ordinates of all ($\sim 10^{23}$) particles and then deducing the value of observables from that wavefunction. There is another approach to many-body problems that we often encounter. In this approach we reduce the number of co-ordinates right from the beginning by trying to describe the system with an order parameter or a density functional.

In particular the "order parameter" is simply a function of co-ordinates that has a finite value (may or may not be spatially constant) in the interesting phase and is zero in the normal phase. Like the wavefunction, the order parameter is not an observable, but can be used to calculate them. The procedure is typically :

- We somehow make a physically motivated guess of the form of free energy, that is *functionally* dependent on the order parameter. Recall the difference between a function and a functional.
- This is just like writing down the Lagrangian or Hamiltonian of a system. The only "proof" that we are correct is that the predictions made on the basis of this turn out to be correct.

In this case Landau and Ginzburg proposed that for a superconductor with an order parameter $\psi(r)$ the free energy density can be written as ($T < T_c$):

$$F[\psi(r)] = F_n + \alpha|\psi(r)|^2 + \beta|\psi(r)|^4 + \frac{1}{2m} |(-i\hbar\nabla + q\mathbf{A})\psi(r)|^2 + \frac{B^2}{2\mu_0} \quad (7.29)$$

Now we have to minimise it in the same way one minimises a lagrangian density. Here the field variable is $\psi(r)$. The answer we get looks very similar to the form of the Schrodinger equation

$$\begin{aligned} \alpha\psi(r) + \beta|\psi(r)|^2\psi(r) + \frac{1}{2m}(-i\hbar\nabla + q\mathbf{A})^2\psi(r) &= 0 \\ \therefore \left[\frac{1}{2m}(-i\hbar\nabla + q\mathbf{A})^2 + \beta|\psi(r)|^2 \right] \psi(r) &= -\alpha\psi(r) \end{aligned} \quad (7.30)$$

Associated with this is a *particle current* density

$$\mathbf{j} = \frac{1}{2} \left[\left(\frac{-i\hbar\nabla - q\mathbf{A}}{m} \psi(r) \right)^* \psi + \psi^* \left(\frac{-i\hbar\nabla - q\mathbf{A}}{m} \psi(r) \right) \right] \quad (7.31)$$

There is no reason a-priori reason to believe that the charge and mass that appear in the equation would be the free electron charge and mass. Now with this expression for current density let us take another look at the Meisner effect. Suppose there is no magnetic field ($\mathbf{A} = 0$) initially. Then we must have the current to be zero everywhere.

$$\begin{aligned} \mathbf{j}(\mathbf{A} = 0) &= \frac{q}{2} \left[\left(\frac{-i\hbar\nabla}{m} \psi(r) \right)^* \psi + \psi^* \left(\frac{-i\hbar\nabla}{m} \psi(r) \right) \right] \\ &= 0 \end{aligned} \quad (7.32)$$

If we assume that $\psi(r)$ is not affected when a magnetic field is turned on, then

$$\begin{aligned} \mathbf{j}(\mathbf{A}) &= \frac{q}{2} \left[\left(\frac{-i\hbar\nabla - q\mathbf{A}}{m} \psi(r) \right)^* \psi + \psi^* \left(\frac{-i\hbar\nabla - q\mathbf{A}}{m} \psi(r) \right) \right] \\ &= \frac{q}{2} \left[\left(\frac{-i\hbar\nabla}{m} \psi(r) \right)^* \psi + \psi^* \left(\frac{-i\hbar\nabla}{m} \psi(r) \right) \right] - \frac{q^2}{m} \psi(r)^* \psi(r) \mathbf{A} \\ &= 0 - \frac{q^2}{m} \psi(r)^* \psi(r) \mathbf{A} \end{aligned} \quad (7.33)$$

Thus the "rigidity" of the wavefunction against a magnetic field can lead to the Meissner effect. There is a little more that it points to. From perturbation theory we know that the change in the ground state wavefunction due to perturbation can be written as :

$$\psi(r, \mathbf{A}) = \psi(r, 0) + \sum_{n \neq 0} \frac{\langle n | H_{pert} | 0 \rangle}{E_n - E_0} |n\rangle \quad (7.34)$$

If there is a considerable gap between the ground state and the first excited state then the ground state wavefunction may change only very little. The rigidity of the wavefunction is an indicator of the existence of a gap, not a proof of it.

7.2.1 Another way of writing the current density: flux quantization

Any wavefunction can be written as

$$\psi(r) = \sqrt{\rho(r)}e^{i\theta(r)} \quad (7.35)$$

It is left as an exercise to show that the equation 7.31 then looks much simpler, written in terms of the phase gradient

$$\mathbf{j} = \frac{q\hbar}{m}\rho\left(\nabla\theta - \frac{q}{\hbar}\mathbf{A}\right) \quad (7.36)$$

A striking prediction now results from the following consideration.

- Say $T > T_c$. We take a ring of a superconducting material and apply a magnetic field along the axis.
- We now cool the ring to $T < T_c$. The field is kept the same during cooldown.
- Clearly surface currents would be set up which cancel the magnetic field in the interior exactly. This is the Meisner effect.
- We now switch off the external field. The surface currents cannot decay and the field created by them would give rise to a magnetic moment.
- What can we say about the field that now threads the hollow of the ring?

Consider a loop deep inside the body of the ring and a closed loop enclosing the hollow of the ring. Since $\mathbf{j} = 0$, we must have

$$\begin{aligned} \nabla\theta &= \frac{q}{\hbar}\mathbf{A} \\ \therefore \oint \nabla\theta \cdot d\mathbf{l} &= \frac{q}{\hbar} \oint \mathbf{A} \cdot d\mathbf{l} \\ \therefore 2n\pi &= \frac{q}{\hbar} \int_{\text{area}} \mathbf{B} \cdot d\mathbf{s} \\ \therefore 2n\pi &= \frac{q}{\hbar} \Phi_{\text{hole}} \end{aligned} \quad (7.37)$$

Unless the loop integral of the phase is $2n\pi$ the wavefunction will not be single valued. This forces the flux through the loop to be quantised. So far we have not said what the charge q is. But a measurement can now tell us this information. Experimentally we find $q = 2e$, a striking proof of pairing. The quantization of flux happens in units of

$$\begin{aligned} \Phi_0 &= \frac{h}{2e} \\ &= 2.07 \times 10^{-15} \text{ Weber} \\ &= 2.07 \times 10^{-7} \text{ gauss.cm}^2 \end{aligned} \quad (7.38)$$

Another macroscopic wavefunction

We mention in the passing another example where a single wavefunction can be used to describe a macroscopic system. This describes a Bose-Einstein condensate. This phenomena does have close links with superconductivity but is not exactly the same. The equation is called the Gross-Pitaevskii equation.

For N bosons interacting via a contact potential only, we can write

$$H = \sum_{i=1}^N \left(\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}_i) \right) + \sum_{i < j} \frac{4\pi\hbar^2 a_s}{m} \delta(\mathbf{r}_i - \mathbf{r}_j) \quad (7.39)$$

The number of particles is conserved, and we can define a chemical potential μ . Then the system can be described with a single co-ordinate wavefunction that resembles the Schrodinger equation with non-linear terms.

$$\left(\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}) + g|\psi(\mathbf{r})|^2 \right) \psi(\mathbf{r}) = \mu\psi(\mathbf{r}) \quad (7.40)$$

where $V(\mathbf{r})$ is the external potential (like the potential of a trap for cold atoms),

$$g = \frac{4\pi\hbar^2 a_s}{m} \quad (7.41)$$

describes the strength of the inter particle interaction. The normalisation is such that

$$\int |\psi(\mathbf{r})|^2 dV = N \quad (7.42)$$

Notice the similarity with the GL equation. The similarity arises because the Free energy of these systems can be (successfully) described with similar looking functional expressions. Minimisation of that w.r.t. the wavefunction leads to the non-linear Schrodinger equation.

7.3 The Cooper pair problem

See the brief introduction to N-particle systems at this point.

The key insight so far has been that a pairing is somehow involved. But pairing requires an attractive interaction. even if we can find a source of lattice deformation mediated attraction between two electrons, a crucial problem remains. In 3D one requires a minimum strength of a potential before a bound state can occur. This argument was important when we analysed the Mott transition. What Cooper showed, was that under certain conditions (that requires the presence of a filled Fermi sea) there can be paired states of two electrons with arbitrarily weak interaction between them. This is a crucial step and requires us to work with the anti-symmetrized wavefunction of two electrons with $k > k_F$. We write

$$\left(\frac{\mathbf{p}_1^2}{2m} + \frac{\mathbf{p}_2^2}{2m} + V(\mathbf{r}_1 - \mathbf{r}_2) \right) \Psi(\mathbf{r}_1\sigma_1, \mathbf{r}_2\sigma_2) = E\Psi(\mathbf{r}_1\sigma_1, \mathbf{r}_2\sigma_2) \quad (7.43)$$

The normalised single particle states are like

$$\phi_{\mathbf{k},\alpha}(\mathbf{r}) = \frac{1}{\sqrt{V}} e^{i\mathbf{k} \cdot \mathbf{r}} \alpha \quad (7.44)$$

where α or β denote the spin functions. Ψ is a Slater determinant built out of these states. Now recall that of the Slater dets that can be formed out of states with zero total momentum

- The $S = 0$ state has symmetric spatial part and antisymmetric spin part.
- The $S = 1$ state has antisymmetric spatial part and symmetric spin part.

So the most general Ψ with $S = 0$ and $K = 0$ can be written as

$$\Psi = \sum_{\mathbf{k}} g(\mathbf{k}) \frac{1}{V} e^{i\mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)} \frac{1}{\sqrt{2}} [\alpha_1\beta_2 - \alpha_2\beta_1] \quad (7.45)$$

with $g(\mathbf{k}) = g(-\mathbf{k})$

We could have written out a general $S_z = -1$ state as

$$\Psi = \sum_{\mathbf{k}} g(\mathbf{k}) \frac{1}{V} e^{i\mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)} \beta_1\beta_2 \quad (7.46)$$

with $g(\mathbf{k}) = -g(-\mathbf{k})$.

Now we force the two electrons to stay above the Fermi sea, by *requiring* that

$$g(\mathbf{k}) = 0 \quad \text{for } k < k_F \quad (7.47)$$

We also want an attractive potential between electrons:

$$V(\mathbf{r}_1 - \mathbf{r}_2) < 0 \quad (7.48)$$

if both \mathbf{k}_1 and \mathbf{k}_2 lies within a shell of width $\hbar\omega_D$ just above E_F .

We must be able to solve for $g(\mathbf{k})$ to be able to make any more progress. Convert eqn 7.43 to an integral problem by

- Insert the general wavefunction eqn 7.45 into equation 7.43.
- left multiply with $\frac{1}{\sqrt{V}}e^{i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)}$
- Integrate over \mathbf{r}_1 and \mathbf{r}_2 and sum over the spin indices

We get

$$(2\varepsilon_k - E)g(\mathbf{k}) + \sum_{\mathbf{k}'} V_{\mathbf{k}\mathbf{k}'}g(\mathbf{k}') = 0 \quad (7.49)$$

where $\varepsilon_k = \hbar^2 k^2 / 2m$ are the single particle energy levels. We also have

$$V_{\mathbf{k}\mathbf{k}'} = \int \int \frac{1}{V} e^{-i\mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)} V(\mathbf{r}_1 - \mathbf{r}_2) \frac{1}{V} e^{i\mathbf{k}' \cdot (\mathbf{r}_1 - \mathbf{r}_2)} d^3\mathbf{r}_1 d^3\mathbf{r}_2 \quad (7.50)$$

$$E_F < \varepsilon_{\mathbf{k}}, \varepsilon_{\mathbf{k}'} < E_F + \hbar\omega_D \quad (7.51)$$

Further

$$V_{\mathbf{k}\mathbf{k}'} = -\frac{V_0}{N} \quad \text{if } \varepsilon_F < \varepsilon_{\mathbf{k}}, \varepsilon_{\mathbf{k}'} < \varepsilon_F + \hbar\omega_D \quad (7.52)$$

$$= 0 \quad \text{otherwise} \quad (7.53)$$

With these simplifications we get:

$$(2\varepsilon_{\mathbf{k}} - E)g(\mathbf{k}) - \frac{V_0}{N} \sum_{\mathbf{k}'} g(\mathbf{k}') = 0 \quad (7.54)$$

Notice that if $g(\mathbf{k})$ is antisymmetric in \mathbf{k} then the sum will vanish. And the solutions would simply be the $\varepsilon_{\mathbf{k}}$. So only the energies of the $S_z = 0$ states can be modified, though we didn't put any spin dependent interaction by hand. We can also see that if we force the spins to align with a magnetic field, then this new state cannot be favoured.

Now we write the solution for $g(\mathbf{k})$, then sum both terms of eqn 7.54 over \mathbf{k} to get a solution for E . It is left as an exercise to complete the algebra. Then we convert the sum over \mathbf{k} to an integral over ε following the usual process and introducing the normal density of states for a metal. We get

$$\frac{V_0}{N} \sum_{\mathbf{k}} \frac{1}{2\varepsilon_{\mathbf{k}} - E} = 1 \quad (7.55)$$

$$\frac{V_0}{N} \int_{\varepsilon_F}^{\varepsilon_F + \hbar\omega_D} D(\varepsilon) \frac{1}{2\varepsilon - E} d\varepsilon = 1 \quad (7.56)$$

Now the question is does this give solutions with $E > 0$? According to the way we have written the equations, the binding energy would be $2\varepsilon_f - E$.

$$\frac{V_0}{N} \int_{\varepsilon_F}^{\varepsilon_F + \hbar\omega_D} D(\varepsilon) \frac{1}{2\varepsilon - E} d\varepsilon = 1$$

$$\therefore \frac{1}{2} V_0 \frac{D(\varepsilon_F)}{N} \ln \frac{2\varepsilon_F + 2\hbar\omega_D - E}{2\varepsilon_F - E} = 1 \quad (7.57)$$

$$\therefore 2\varepsilon_F - E = \hbar\omega_D \frac{e^{-1/V_0 n(\varepsilon_F)}}{\sinh 1/V_0 n(\varepsilon_F)} \quad (7.58)$$

$$\approx 2\hbar\omega_D e^{-2/V_0 n(\varepsilon_F)} \quad \text{if } V_0 n(\varepsilon_F) \ll 1 \quad (7.59)$$

The result tells us

- Even if V_0 is very small we are going to get a solution. The blocking of all the states below ε_f have removed the constraint of a minimum strength of potential, even in 3D.
- The binding energy is not an analytic function of V_0 - hence we could not have got this result by expanding a complicated potential and treating it perturbatively.
- If there is a large V_0 , arising from electron-lattice interaction, it would favour the bound state. In a loose sense, it tells us why very good metals (less electron-lattice coupling) like gold, copper etc. are not found to be superconducting.

7.3.1 Many pairs together

It can be shown that the size of a pair, is much large than the mean separation of the electrons in the metal. Thus if such a state at all forms, many pairs would be strongly overlapping and we must have a way of treating them all at once. Here we must use the many body states, using the creation and annihilation operators we have introduced (see the appendix).

The normal metal and a single Cooper pair

The normal metal (Fermi sea at $T = 0$) with N electrons, can be described by:

$$|\Psi_N\rangle = \prod_{\mathbf{k}}^{k < k_F} c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}\downarrow}^\dagger |0\rangle \quad (7.60)$$

We can then write a single Cooper pair as

$$|\Psi_{Cooper\ Pair}\rangle = \sum_{\mathbf{k}} g(\mathbf{k}) c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}\downarrow}^\dagger |\Psi_N\rangle \quad (7.61)$$

We did not need to specify $k > k_F$ in the last expression - can you see why?

The Hamiltonian

We write the hamiltonian using the creation/annihilation operators

$$H = \sum_{\mathbf{k}} \varepsilon_{\mathbf{k}} \left(c_{\mathbf{k}\uparrow}^\dagger c_{\mathbf{k}\uparrow} + c_{-\mathbf{k}\downarrow}^\dagger c_{-\mathbf{k}\downarrow} \right) + \sum_{\mathbf{k}\mathbf{k}'} V_{\mathbf{k}\mathbf{k}'} c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}\downarrow}^\dagger c_{-\mathbf{k}'\downarrow} c_{\mathbf{k}'\uparrow} \quad (7.62)$$

The Bardeen Cooper Schrieffer (BCS) wavefunction

A variational wavefunction was written by BCS and then minimised. The wavefunction is not a fixed N wavefunction. It is written with an infinite number of variational parameters $u_{\mathbf{k}}, v_{\mathbf{k}}$ as

$$|\Psi_{BCS}\rangle = \prod_{\mathbf{k}} \left(u_{\mathbf{k}} + v_{\mathbf{k}} c_{\mathbf{k}\uparrow}^{\dagger} c_{-\mathbf{k}\downarrow}^{\dagger} \right) |0\rangle \quad (7.63)$$

We state the following without proof, but you can work it out...

- $u_{\mathbf{k}}, v_{\mathbf{k}}$ are real and even.
- $u_{\mathbf{k}}^2 + v_{\mathbf{k}}^2 = 1$ ensures that the state $|\Psi_{BCS}\rangle$ is normalised to 1.
- $v_{\mathbf{k}}$ represents the probability that the state $(\mathbf{k} \uparrow, -\mathbf{k} \downarrow)$ is occupied
- It follows that $u_{\mathbf{k}}$ will represent the probability that it is unoccupied.
- The normal metal has

$$\begin{aligned} u_{\mathbf{k}} &= 0 & v_{\mathbf{k}} &= 1 & \text{for } k < k_F \\ u_{\mathbf{k}} &= 1 & v_{\mathbf{k}} &= 0 & \text{for } k > k_F \end{aligned} \quad (7.64)$$

The BCS state does not have a fixed number of particles. The number operator is

$$\hat{N} = \sum_{\mathbf{k}} \left(c_{\mathbf{k}\uparrow}^{\dagger} c_{\mathbf{k}\uparrow} + c_{-\mathbf{k}\downarrow}^{\dagger} c_{-\mathbf{k}\downarrow} \right) \quad (7.65)$$

Now for the superconducting wavefunction we need to

- minimise $\langle \Psi_{BCS} | H | \Psi_{BCS} \rangle$ subject to the constraint $\langle \Psi_{BCS} | \hat{N} | \Psi_{BCS} \rangle = N$.

This clearly is a "Lagrange multiplier" problem. The procedure and the result are now very well known - we will use the final result and skip the procedure. Just note that our variables are $u_{\mathbf{k}}, v_{\mathbf{k}}$ and the multiplier itself - which turns out to be the chemical potential (μ) of the system. μ is fixed by N .

The values of the gap, $u_{\mathbf{k}}$ and $v_{\mathbf{k}}$

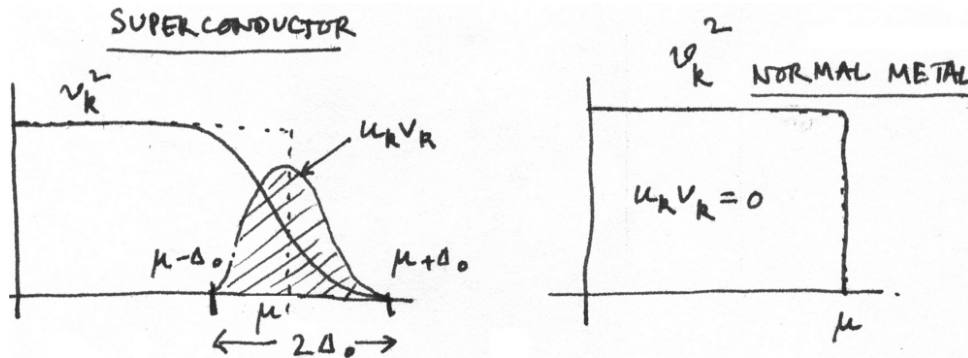


Figure 7.5: The solution of the variational parameters obtained after optimising the BCS Hamiltonian. Notice how the region of interest turns out to be around the Fermi level of the normal metal.

The single particle energy gap is given by

$$\Delta(0) = 2\hbar\omega_D e^{-\frac{1}{V_0 n(\varepsilon_F)}} = 1.76 k_B T_c \quad (7.66)$$

$$\Delta(T) = 3.06 k_B T_c \sqrt{1 - \frac{T}{T_c}} \quad \text{for } T \approx T_c \quad (7.67)$$

The solution for the variational parameters are (see fig 7.5

$$\begin{aligned} u_{\mathbf{k}}^2 &= \frac{1}{2} \left[1 + \frac{\varepsilon_{\mathbf{k}} - \mu}{\sqrt{(\varepsilon_{\mathbf{k}} - \mu)^2 + \Delta^2}} \right] \\ v_{\mathbf{k}}^2 &= \frac{1}{2} \left[1 - \frac{\varepsilon_{\mathbf{k}} - \mu}{\sqrt{(\varepsilon_{\mathbf{k}} - \mu)^2 + \Delta^2}} \right] \end{aligned} \quad (7.68)$$

The expectation value of the pair creation operator distinguishes the normal state from the superconducting state

$$\langle \Psi_{BCS} | c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}\downarrow}^\dagger | \Psi_{BCS} \rangle = u_{\mathbf{k}} v_{\mathbf{k}} = \frac{1}{2} \frac{\Delta(T)}{\sqrt{(\varepsilon_{\mathbf{k}} - \mu)^2 + \Delta(T)^2}} \quad (7.69)$$

This quantity is zero in the normal state, but in the superconducting state it is different in a shell of width Δ , around the chemical potential. From the free electron dispersion this width in energy translates to a width in k

$$\delta E = \Delta = \frac{\hbar k_F}{m} \delta k = \hbar v_F \delta k \quad (7.70)$$

Thus the pair wavefunction will have a spatial width (called the coherence length ξ_0) of the order

$$\xi_0 = \frac{\hbar^2 k_F}{\pi m \Delta} \quad (7.71)$$

In superconductors like Pb, Sn, Nb *etc.* $\xi_0 \sim 1000 \text{\AA}$. The spatial extent of a Cooper pair is thus very large compared to the mean separation between the pairs. The states are thus strongly overlapping, they could not have been treated in isolation.

Chapter 8

Josephson junctions

The junction between a superconductor and a normal metal can show qualitatively different conductance characteristics depending on how strong the barrier is. We just show a picture of this phenomena but would not analyse it further.

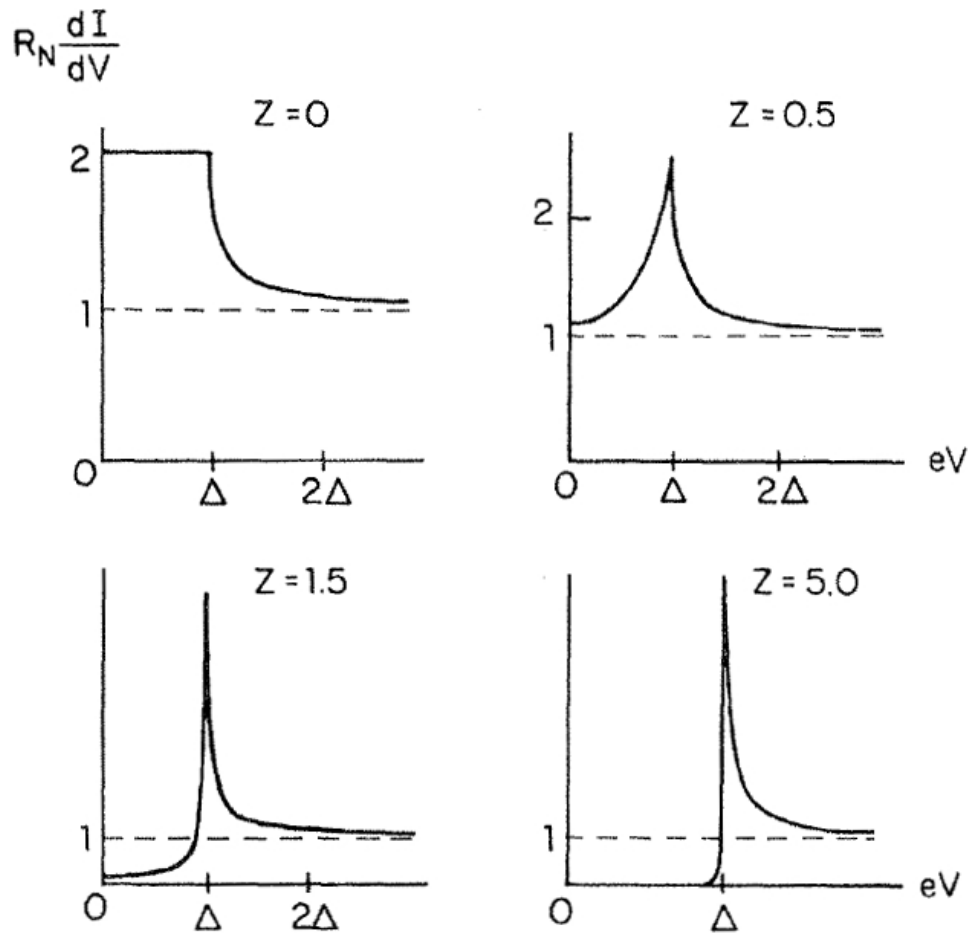


Figure 8.1: Differential conductance vs voltage for various barrier strengths Z at $T=0$. This quantity is proportional to the transmission coefficient for electric current for particles at $E=eV$. The figure is taken from "Transition from metallic to tunneling regimes in superconducting microconstrictions: Excess current, charge imbalance, and supercurrent conversion", by G. E. Blonder, M. Tinkham, and T. M. Klapwijk, *Physical Review B*, **25**, 4515 (1982). Notice how the behaviour for subgap bias is markedly different depending on how strong the barrier is.

Now what happens in a superconductor-superconductor (SS) junction? The barrier separating the

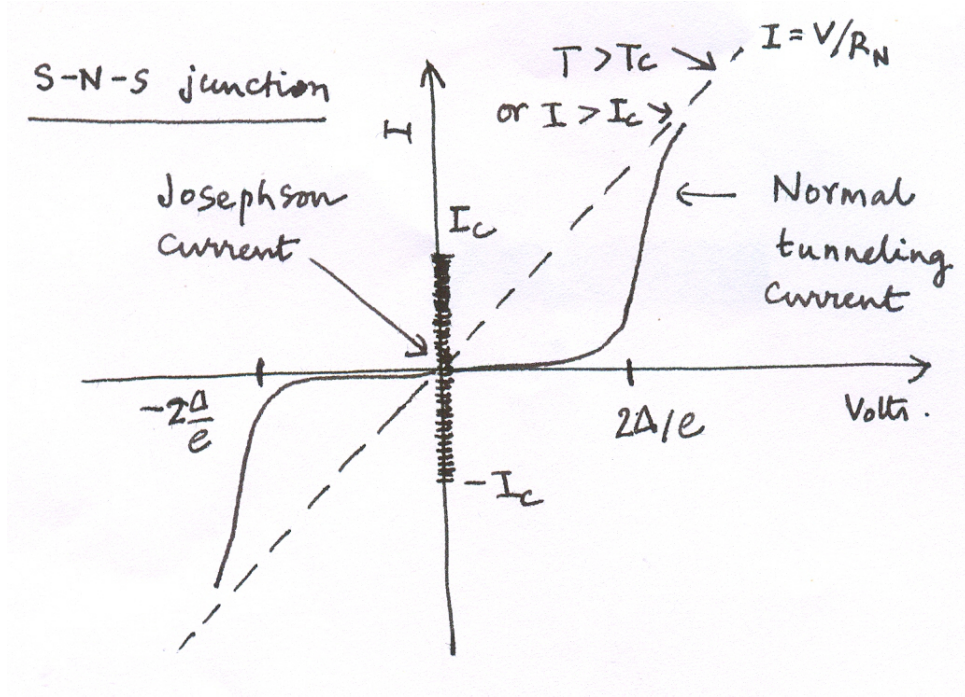


Figure 8.2: The Josephson current, normal state resistance and the normal (tunnel) current in the junction.

two regions can be thin or thick. One may guess that there will be two components of the current. Single electrons may tunnel if the voltage excitation is high enough. Also Cooper pairs may tunnel at voltage biases below the gap. Josephson junctions are essentially S-N-S or S-I-S junctions where a significant amount of the current is carried by the Cooper pairs.

We want to be able to calculate the I-V characteristics. But how do we setup the problem? We will give an overview of how to do it using the macroscopic (Landau Ginzburg order parameter) and microscopic (BCS wavefunction and tunnelling) viewpoints. It is easier to derive the basic result using the Landau-Ginzburg method. We do that first. After this we will drive the same using the microscopic picture.

Josephson junction from the Ginzburg-Landau order parameter

See the figure 8.3. The order parameters from the left and right decay in the barrier. So that in the barrier we can write, as a reasonable guess in the barrier region

$$\psi(z) = \psi_1 e^{-\beta z} + \psi_2 e^{\beta(z-b)} \quad (8.1)$$

where β is the damping in the barrier region. But this implies a current of the form

$$\begin{aligned} J(z) &= -\frac{i\hbar e}{2m} \left[\psi^* \frac{d\psi}{dz} - \psi \frac{d\psi^*}{dz} \right] \\ &= -\frac{i\hbar e}{m} \beta e^{-b\beta} (\psi_1^* \psi_2 - \psi_2^* \psi_1) \\ &= \frac{2\hbar e}{m} \beta e^{-b\beta} |\psi_1| |\psi_2| \sin(\theta_1 - \theta_2) \end{aligned} \quad (8.2)$$

We have assumed, via the expression for current that there is no magnetic field. We will have to put that in later. Now consider a voltage applied across the junction. The relative energy difference between the Cooper pairs must be $2eV$. In analogy with the time dependent Schrodinger equation,

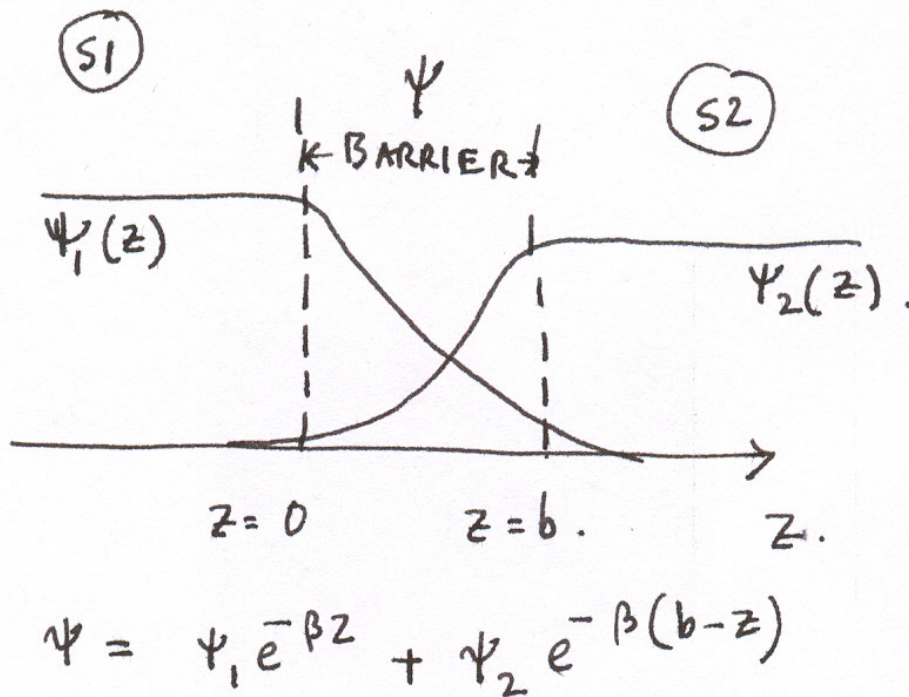


Figure 8.3: The decay of the order parameter in the barrier.

we expect that the time variation of the relative phase $\phi = \theta_1 - \theta_2$ to be

$$I = I_C \sin \phi \quad (8.3)$$

$$\frac{d\phi}{dt} = \frac{2eV}{\hbar} \quad (8.4)$$

An alternative derivation (see the Feynman lectures) is set as a problem.

The macroscopic approach tells us that the decay of the order parameter into the barrier lets it connect or couple with the wavefunction on the other side. Hence the particle exchange is possible. But it does not tell us things like

1. The relation of the critical current and the gap size.
2. Relative values of the normal to supercurrent.

The microscopic view : using BCS

Here we give the outline of a derivation given by deGennes and Ferrel. We do not calculate every matrix element on the way - but will state the final result in some cases. Originally the striking prediction of a supercurrent at zero voltage across the junction was made by Josephson. The hamiltonian for the system ($S + S'$) is given by the uncoupled and the "tunneling" parts

$$H = H_{SS'} + H_T \quad (8.5)$$

$$H_T = \sum_{kl} \left(T_{kl} c_{kS}^\dagger c_{lS'} + T_{kl}^\dagger c_{lS'}^\dagger c_{kS} \right) \quad (8.6)$$

Note that the c and c^\dagger are electron creation and annihilation operators not pair creator/annihilator. Thus to transfer a pair (two electrons) we would have to look for a second order process - ensuring

two actions by the operator. Let us write a state with ν pairs in S and $N - \nu$ pairs in S' . We assume that altogether there are $2N$ electrons in $S + S'$. Let:

$$\Psi_\nu = \Phi_{2(N-\nu)}^{S'} \Phi_{2\nu}^S \quad (8.7)$$

$$H_{SS'} \Psi_\nu = E_\nu \Psi_\nu \quad (8.8)$$

Tunneling has been ignored till now. What is the cost of transferring two electrons from one side to the other? by definition:

$$E_\nu - E_{\nu-1} = 2(E_F^S - E_F^{S'}) \quad (8.9)$$

This difference can be fixed by an external battery. We now need to calculate the second order connection between the states with one pair less (or more)

$$J_0 = \sum_{kl, k'l'} \langle \nu + 1 | T_{kl} c_{kS}^\dagger c_{lS'} | I \rangle \frac{1}{E - E_I} \langle I | T_{k'l'} c_{k'S}^\dagger c_{l'S'} | \nu \rangle \quad (8.10)$$

- The intermediate state $|I\rangle$ contains $2\nu + 1$ electrons on side S and $2(N - \nu) - 1$ electrons on the side S' where the extra electron has an wavevector k' .
- Then the second action of H_T on $|I\rangle$ creates another electron with wavevector k .
- The final state onto which this is projected has $\nu + 1$ pairs on side S .
- This would imply that we only need to keep terms $k' = -k$ and $l' = -l$.

Using the BCS wavefunction and the excited states (with one extra unpaired particle) we can show that

$$J_0 = -4 \sum_{kl} |T_{kl}|^2 \frac{u_k v_k u_l v_l}{\varepsilon_k + \varepsilon_l} \quad (8.11)$$

Thus to second order in H_T there is a connection between the states Ψ_ν and $\Psi_{\nu+1}$ and we can write:

$$H \Psi_\nu = E \Psi_\nu + J_0 (\Psi_{\nu+1} + \Psi_{\nu-1}) \quad (8.12)$$

The eigenfunction of H would then be a linear combination of many Ψ_ν . In a very similar way, the tight binding wavefunctions are linear combinations of wavefunctions centered at many sites. In fact the mathematical similarity of this problem with the nearest neighbour tight binding is striking. The index ν takes the place of the site index. We use it in the following way, by introducing a variable k , such that

•

$$\Psi_k = \sum_\nu e^{ik\nu} \Psi_\nu \quad (8.13)$$

The variables k and ν appear in the same fashion as x and p .

- The energy eigenvalues of this "tight binding" hamiltonian must be

$$E(k) = E_0 + 2J_0 \cos k \quad (8.14)$$

- Then the "wave packet" formed with states built around a certain k value will move with a group velocity

$$\frac{d}{dt} \langle \hbar k \rangle = -2J_0 \sin k \quad (8.15)$$

- It follows that the current is

$$I = -2e \frac{2J_0}{\hbar} \sin k \quad (8.16)$$

- If there is a chemical potential difference between the two sides, then we can apply the analogy of an electric field acting on the "Bloch electrons"

$$\frac{d}{dt}\langle\hbar k\rangle = 2eV \quad (8.17)$$

- Finally, notice that all these effects will vanish if the product $u_k v_k = 0$. Hence none of these happen in the normal state.

8.1 The RCSJ model

p-n junctions and heterointerfaces form the fundamental building blocks of semiconductor electronics, Josephson junctions are the basic components of superconducting circuits. We need to understand its current-voltage characteristics.

The equations governing a single junction shunted by a capacitor and a resistance, are discussed in detail in several textbooks [Tinkham 1996]. This is the model that we will use in this chapter. The current flowing through the junction has three components- the "supercurrent" (I_s) caused by the phase difference (ϕ) across the junction, the current through the resistance (R) caused by the potential difference (V) and the current through the capacitor (C). The externally impressed voltage causes ϕ to change with time :

$$I_s = I_c \sin \phi \quad (8.18)$$

$$\frac{d\phi}{dt} = \frac{2eV}{\hbar} \quad (8.19)$$

Using 8.18&8.19 show that the coupling energy between the two superconductors forming the junction is given by

$$\int I_s V dt = \text{const.} - E_J \cos \Delta\phi \quad \text{where} \quad E_J \equiv \frac{\hbar I_c}{2e} \quad (8.20)$$

The total current should be:

$$I = \frac{V}{R} + I_c \sin \phi + C \frac{dV}{dt} \quad (8.21)$$

By introducing the variables

$$\begin{aligned} i &= I/I_c \\ \omega &= \frac{2e}{\hbar} I_c R \\ \tau &= \omega t \\ \beta &= \omega RC \end{aligned}$$

we get

$$\beta \ddot{\phi} + \dot{\phi} + \sin \phi = i \quad (8.22)$$

The parameter β is called the "Stewart-McCumber" parameter. The dots denote differentiation w.r.t to τ , β is the single parameter that characterises the behaviour of the junction. The fast variations of the phase occur with a frequency $\omega \sim 10^{12}$ Hz, what the voltmeter measures is an average value of $\dot{\phi}$ over several cycles. If we sweep the current through a junction then the voltage jumps from zero to IR as I crosses I_c . If the current is then reduced below I_c , the voltage reduces to zero at a lower value of current called the retrapping current (I_r). The ratio I_r/I_c depends only on β . Figure 8.4 shows numerically calculated I - V traces of a single junction for various values of β . The physical reasons behind the hysteretic behaviour are described in several textbooks, (*e.g.* see Tinkham's book). We

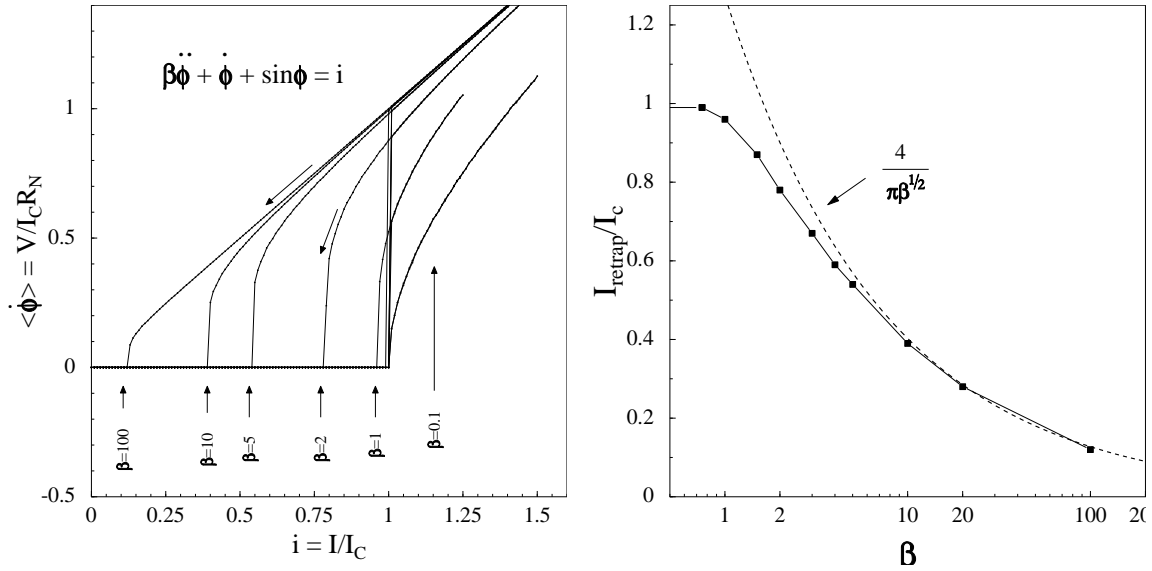


Figure 8.4: Numerically calculated I - V of a Josephson-junction for various values of β . The width of the hysteresis increases with increasing β . From the observed ratio of the retrapping current and critical current, it is possible to infer the value of β . For $\beta \approx 10$ or more, the curve of I_r/I_c to β can be approximated by $I_r = 4I_c/(\pi\sqrt{\beta})$.

describe the *tilted washboard* model.

PROBLEM : Another way of scaling the junction equation is as follows. Introduce the variables:

$$\begin{aligned} i &= \frac{I}{I_c} \\ \omega_p &= \sqrt{\frac{2eI_c}{\hbar C}} \\ Q &= \omega_p RC \\ \tau &= \omega_p t \end{aligned}$$

Show that this leads to the equation

$$\frac{d^2\phi}{d\tau^2} + \frac{1}{Q} \frac{d\phi}{d\tau} + \sin\phi = \frac{I}{I_c} \quad (8.23)$$

This emphasizes the dissipative term (using the parameter Q), whereas eqn. 8.22 emphasized a mass like inertia term. Notice that the scaling of the time variable is different in the two cases. From the definitions show that:

$$\beta = Q^2 \quad (8.24)$$

Obviously a small Q means that the second derivative (inertia) term is less important, large β means that the dissipative part (comes from the resistance) is less important.

8.1.1 The titled washboard

Both eqns. 8.22 and 8.23 admit a mechanical analogy. We have one term which looks like mass \times acceleration, one term like frictional/viscous dissipation and a force like term derivable from a position dependent

potential

$$u(\phi) = -i\phi - \cos \phi \quad \text{scaled units} \quad (8.25)$$

$$U(\phi) = -E_J \cos \phi - \frac{\hbar I}{2e} \phi \quad \text{unscaled units} \quad (8.26)$$

- This potential has an overall slope given by the drive current and an oscillating part.
- If $I < I_c$ then a solution with $\phi = \text{constant}$ is possible. This is the part where there is current flow without a voltage drop. Let us start driving the RCSJ with a constant current source, starting from $I = 0$.
- now we use the mechanical analogy. As the potential tilts more and more a particle trapped at one of the many minimas of the "titled washboard" will spill out and start moving. ϕ can no longer be constant and there will be a voltage across the junction.
- The equations if numerically solved will give a rapidly oscillating $\phi(t)$. This fast oscillations are not important. The voltmeter will measured the time averaged ϕ . It can be shown that in this (normal) state the $I - V$ behaviour of the junction is nearly that of a resistance.
- Now let us start decreasing the current. Will it drop back to zero right at I_c ? This depends on the value of β (or Q , depending on how we have scaled the equations).
- First consider the case of small inertia or large dissipation. Under these circumstances, the $\ddot{\phi}$ can be ignored. The equation 8.23 is then exactly integrable.

$$\frac{d\phi}{dt} = \frac{2eI_c R}{\hbar} \left(\frac{I}{I_c} - \sin \phi \right) \quad (8.27)$$

Now calculate how much time it would take for the phase to change by 2π . Let us say this time is T .

- The average voltage measured by a voltmeter across the junction will then by

$$\frac{2e}{\hbar} \langle V \rangle = \frac{2\pi}{T} \quad (8.28)$$

It is left as a problem to show that for $I > I_c$

$$\langle V \rangle = R \sqrt{I^2 - I_c^2} \quad (8.29)$$

- Thus for small β or small Q , the equation is approximately solvable in closed form. This is called the OVERDAMPED case. Physically it means that the capacitance of the junction can be ignored. In electrical circuits capacitors and inductors often behave like mass, because some charge or current in those elements can store energy, very much like a mass "stores" some kinetic energy when it moves.
- See the curve for $\beta = 0.1$ in figure 8.4 and try to relate it with what we just calculated. There is no hysteresis in this limit.
- This is important to note that the jump between the zero voltage and resistive states can occur in about a *picosecond*. Josephson junction is the fastest known switch.
- The next point is to understand what happens when the capacitance and resistance must be simultaneously taken into account. We will continue to assume that we are at $T = 0$. This assumption has entered in a slightly subtle way in our analysis. We have ignored any thermal fluctuation that can round off the transition or cause the "particle" to escape from the bottom of one cycle of the titled washboard.

- Again we begin in the same way and start increasing the current from $I = 0$. At I_c the current jumps to the normal state and nearly follows the normal state line. Then we start decreasing the current. This is equivalent to starting to reduce the tilt of the washboard and try to trap the particle again. But now that there is a mass term the particle will not stop at the same tilt that set it running.
- This happens because the acceleration due to the slope of the washboard must be less than the dissipation in every cycle, so that the system takes away some energy from the particle in every cycle. Only then will the particle be *retrapped*. The retrapping current, I_r , at which it happens is related to I_c for large β approximately by:

$$\frac{I_r}{I_c} \approx \frac{4}{\pi\sqrt{\beta}} \quad (8.30)$$

hence the $I - V$ in this case ($\beta \gg 1$) can be highly hysteretic. See the fig. 8.4 again.

- Now, how can we understand at what current it will retrap? When we asked whether the particle will start rolling, if placed initially at rest at a local minima. Now we ask whether a particle places at a local maxima with near zero kinetic energy, will be able to reach the next maxima. We need to compare how much energy will be gained due to the tilt and how much will be lost due to dissipation as it advances by 2π .
- The energy fed by the current over one cycle, $\Delta\phi = 2\pi$,

$$\begin{aligned} \int IV dt &= \int I \frac{\hbar}{2e} \frac{d\phi}{dt} dt \\ &= I \frac{\hbar}{2e} \end{aligned} \quad (8.31)$$

The dissipation over one cycle matches this then that will be the borderline (current) below which free running solutions are not possible. It is somewhat non-trivial to deduce the full expression for dissipation per cycle¹ which finally leads to eqn. 8.30.

8.1.2 Two different steady states: role of the initial conditions

The equation 8.23 can be numerically solved with the current set to a value somewhat below $I/I_C = 1$. In fig 8.5 you can see what happens to the solution for $\dot{\phi}$ as a function of τ . Depending on the initial conditions there can be two possible values of the time average $\langle \dot{\phi} \rangle$. One is zero and the other is finite. From a purely mathematical point of view this is the origin of the hysteresis.

8.1.3 Introducing the magnetic field

How would the phase difference across a junction change if a magnetic field is introduced? We again fall back on the idea that the order parameter satisfies a Schrodinger like equation and try to deduce something by analogy. Suppose we know a certain wavefunction, $\psi(r)$ satisfies

$$\left[\frac{\mathbf{p}^2}{2m} + V(\mathbf{r}) \right] \psi = i\hbar \frac{\partial \psi}{\partial t} \quad (8.32)$$

Now if we keep $V(\mathbf{r})$ unchanged and introduce a vector potential $\mathbf{A}(\mathbf{r})$ how will the new wavefn be related to the old one? We make use of the following relation (it is left as a simple exercise to show

¹See: The return of a hysteretic Josephson junction to the zero-voltage state: IV characteristic and quantum retrapping. Y. C. Chen, Matthew P. A. Fisher, and A. J. Leggett *Journal of Applied Physics*, **64**, 3119 (1988); & Current-voltage characteristics of Josephson junctions. W. C. Stewart *Applied Physics Letters*, **12**, 277 (1968)

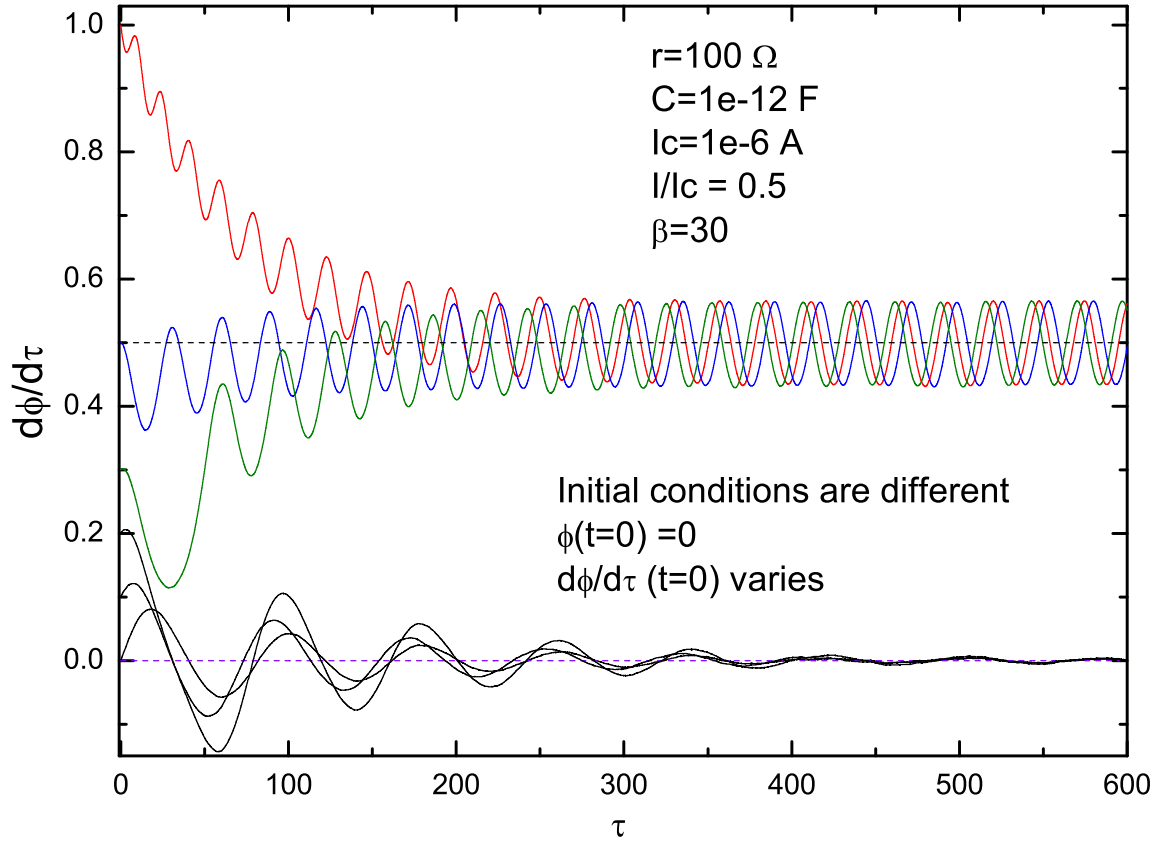


Figure 8.5: The role of initial conditions.

how the extra terms produced by the derivative $\mathbf{p} = -i\hbar\nabla$ cancel)

$$(\mathbf{p} + q\mathbf{A}) \psi e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}} = e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}} \mathbf{p} \psi$$

$$(\mathbf{p} + q\mathbf{A}) (\mathbf{p} + q\mathbf{A}) \psi e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}} = e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}} \mathbf{p}^2 \psi \quad (8.33)$$

$$(8.34)$$

This lets us write:

$$\left[\frac{(\mathbf{p} + q\mathbf{A})^2}{2m} + V(\mathbf{r}) \right] \psi e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}} = i\hbar \frac{\partial}{\partial t} \psi e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}} \quad (8.35)$$

Which implies that only the phase has changed

$$\psi = \sqrt{\rho} e^{i\theta} \rightarrow \sqrt{\rho} e^{i\theta} e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{l}}$$

$$\therefore \theta(\mathbf{r}) \rightarrow \theta(\mathbf{r}) + \frac{q}{\hbar} \int \mathbf{A} \cdot d\mathbf{l} \quad (8.36)$$

Thus the phase difference appearing in the josephson current equation 8.18 needs to be replaced by

$$I_s = I_c \sin \left(\phi + \frac{q}{\hbar} \int \mathbf{A} \cdot d\mathbf{l} \right) \quad (8.37)$$

- The correct choice is $q = -2|e|$ as we know from flux quantization experiments.
- The mass of the electron does not appear here, making the result independent of band structure details.

We are now ready to understand how two josephson junctions connected in parallel and with a magnetic flux (not so large as to kill the superconducting state) threading the area in between will behave.

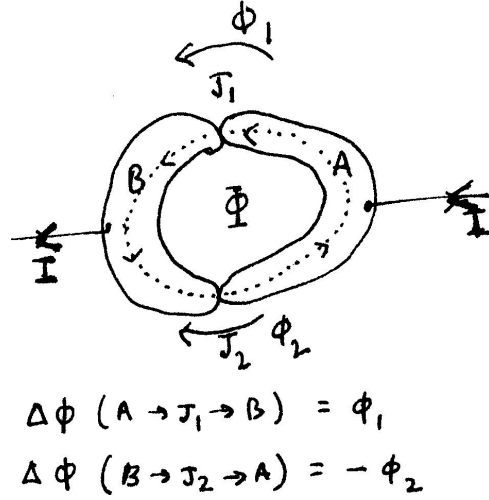


Figure 8.6: Note the sense in which the path is traversed. The sign of the phase difference must be treated correctly. When we write the current equation we want the phase lead of B w.r.t. A. But while taking the integral we travel once from A to B and then from B to A.

8.1.4 Principle of the Superconducting Quantum Interference Device

Let us now analyse what happens when two junctions are in parallel. See fig. 8.6. We take a path along the loop as shown. Along the path the following relation holds:

$$mv_s = \hbar \left(\nabla\phi - \frac{2\pi}{\Phi_0} \mathbf{A} \right) \quad (8.38)$$

where Φ_0 is the flux quantum already defined. The total flux passing through the loop is denoted by Φ .

$$\begin{aligned} \Phi &= \int_{\text{Loop}} \mathbf{A} \cdot d\mathbf{l} \\ &= \int_{\text{Electrodes}} \mathbf{A} \cdot d\mathbf{l} + \int_{J_1} \mathbf{A} \cdot d\mathbf{l} + \int_{J_2} \mathbf{A} \cdot d\mathbf{l} \\ &= \frac{\Phi_0}{2\pi} \int_{\text{Electrodes}} \nabla\phi \cdot d\mathbf{l} + \int_{J_1} \mathbf{A} \cdot d\mathbf{l} + \int_{J_2} \mathbf{A} \cdot d\mathbf{l} \end{aligned} \quad (8.39)$$

We have used eqn 8.38 in the last step. It is important to take the integrals in the correct sense to get the signs of the quantities. In going round the loop the total change in the phase must be $2n\pi$. We write it out explicitly

$$\int_{J_1 \rightarrow B \rightarrow J_2} \nabla\phi \cdot d\mathbf{l} + \Delta\phi(B \rightarrow J_2 \rightarrow A) + \int_{J_2 \rightarrow A \rightarrow J_1} \nabla\phi \cdot d\mathbf{l} + \Delta\phi(A \rightarrow J_1 \rightarrow B) = 2n\pi \quad (8.40)$$

Now it is convenient to write out the change in phase across the junction more simply

$$\begin{aligned} \Delta\phi(A \rightarrow J_1 \rightarrow B) &= \phi_1 \\ \Delta\phi(B \rightarrow J_2 \rightarrow A) &= -\phi_2 \end{aligned} \quad (8.41)$$

So that we consider the change in going from A to B rather than force an anticlockwise sense.

$$\int_{J1 \rightarrow B \rightarrow J2} \nabla \phi \cdot d\mathbf{l} + \int_{J2 \rightarrow A \rightarrow J1} \nabla \phi \cdot d\mathbf{l} + (\phi_1 - \phi_2) = 2n\pi \quad (8.42)$$

The total flux through the loop and the phases are then related as

$$\begin{aligned} \frac{\Phi_0}{2\pi} [2n\pi - (\phi_1 - \phi_2)] + \int_{BJ2A} \mathbf{A} \cdot d\mathbf{l} + \int_{AJ1B} \mathbf{A} \cdot d\mathbf{l} &= \Phi \\ \left(\phi_1 - \frac{2\pi}{\Phi_0} \int_{AJ1B} \mathbf{A} \cdot d\mathbf{l} \right) - \left(\phi_2 - \frac{2\pi}{\Phi_0} \int_{AJ2B} \mathbf{A} \cdot d\mathbf{l} \right) &= 2\pi \frac{\Phi}{\Phi_0} \mod 2\pi \end{aligned} \quad (8.43)$$

In presence of a magnetic field the phase is to be replaced by the combination

$$\gamma = \phi - \frac{2\pi}{\Phi_0} \int_{Path} \mathbf{A} \cdot d\mathbf{l} \quad (8.44)$$

And hence the sum total of the current through the device (two identical junctions in parallel) must be :

$$\begin{aligned} I &= I_c (\sin \gamma_1 + \sin \gamma_2) \\ &= 2I_c \cos \frac{\gamma_1 - \gamma_2}{2} \sin \frac{\gamma_1 + \gamma_2}{2} \\ &= 2I_c \left(\cos 2\pi \frac{\Phi}{\Phi_0} \right) \sin \frac{\gamma_1 + \gamma_2}{2} \end{aligned} \quad (8.45)$$

The maximum supercurrent through the parallel combination of two junctions oscillates with the flux passing through the junction. Since the flux quantum is a small quantity and it is possible to resolve about 10^{-5} of an oscillation, we have a very sensitive detector of flux, or any quantity that can be converted to a flux.

The equivalent circuit of a SQUID

- What the experimenter measures is the total current and the full voltage (I and V in the fig 8.7).
- Since the structure is like a loop we need to take into account its self inductance. This means that the flux linkage due to the current flowing in the branches is not to be neglected. But this current branches out ($I = I_1 + I_2$) and comes together again at the other end of the loop. So we have to find the flux linkage in a situation where the current in the two halves of a loop are different. It is left as a simple exercise.
- Also, we would set the junction parameters in such a way that there is no hysteresis in the junction I-V. From the results of the previous section, this means we should make $\beta < 1$ or make the resistance sufficiently small. This can be done by having a shunt resistance (a thin metal film) across the junction.
- We will show the equations including a new thermal noise current term. This term allows one to model thermal fluctuations that may allow the escape of the phase point from the minimas of the washboard potential. It also has obvious effects on the resolution. We will point out some results but not calculate the details of the noise term.

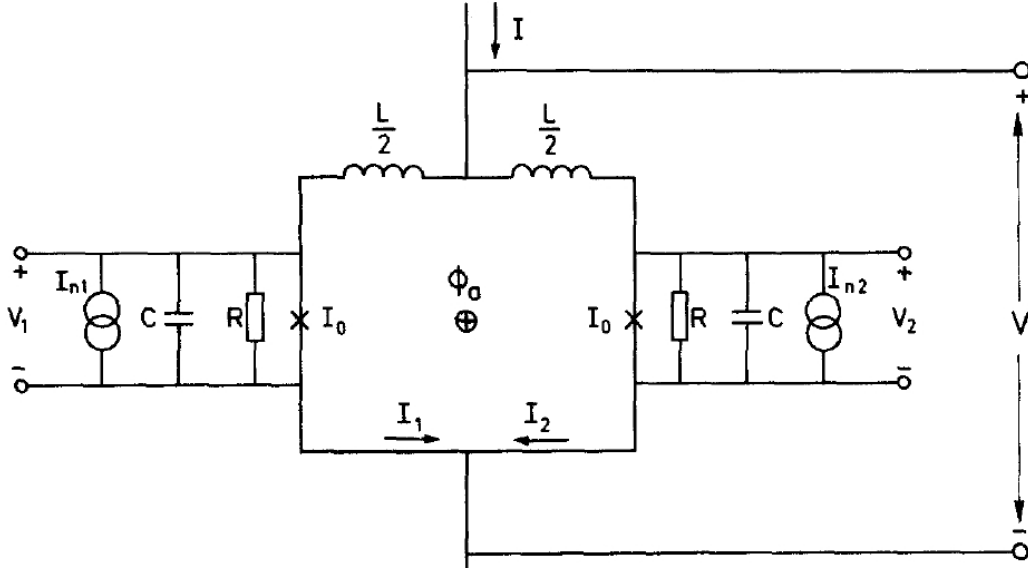


Figure 8.7: The equivalent circuit of a SQUID. We neglect the possible small asymmetry in the R , C , I_c parameters of the junction. It is known that they do not affect the behaviour very significantly. The figure and analysis is taken from "Simulation and optimization of a DC SQUID with finite capacitance", V.J. De Wall and P. Schrijner, *Journal of Low Temperature Physics*, **54**, 215

So the equations of motion are now

$$\begin{aligned} I_1 &= I_c \sin \gamma_1 + I_{1,noise} + \frac{V_1}{R} + C \frac{dV_1}{dt} \\ I_2 &= I_c \sin \gamma_1 + I_{2,noise} + \frac{V_2}{R} + C \frac{dV_2}{dt} \end{aligned} \quad (8.46)$$

$$\begin{aligned} \frac{d\gamma_1}{dt} &= \frac{2eV_1}{\hbar} \\ \frac{d\gamma_2}{dt} &= \frac{2eV_2}{\hbar} \end{aligned} \quad (8.47)$$

The phase difference is determined by the *total* flux through the loop, which is the sum of the applied flux and the self induced flux.

$$\gamma_1 - \gamma_2 = 2\pi \frac{\Phi_t}{\Phi_0} \quad (8.48)$$

$$\Phi_t = \Phi_{ext} + LJ \quad (8.49)$$

$$J = \frac{1}{2}I_2 - \frac{1}{2}I_1 \quad (8.50)$$

From the equivalent circuit, we can relate V to V_1 and V_2

$$\begin{aligned} V &= V_1 - \frac{L}{2} \frac{dJ}{dt} \\ V &= V_2 + \frac{L}{2} \frac{dJ}{dt} \\ \therefore V &= \frac{1}{2}(V_1 + V_2) \end{aligned} \quad (8.51)$$

It is left as an exercise to show that this set of equation can now be scaled like before by introducing

$$\beta = \frac{2eI_c R^2 C}{\hbar} \quad (8.52)$$

$$\beta_L = \frac{2I_c L}{\Phi_0} \quad (8.53)$$

We get after scaling the current (by $2I_c$), voltage (by $I_c R$) and time as before and neglecting the noise term

$$\beta \ddot{\gamma}_1 + \dot{\gamma}_1 + \sin \gamma_1 = \frac{i}{2} - j \quad (8.54)$$

$$\beta \ddot{\gamma}_2 + \dot{\gamma}_2 + \sin \gamma_2 = \frac{i}{2} + j \quad (8.55)$$

$$j = \frac{1}{\pi \beta_L} \left(\gamma_1 - \gamma_2 - 2\pi \frac{\Phi_{ext}}{\Phi_0} \right) \quad (8.56)$$

$$\frac{V}{I_c R} = \frac{\dot{\gamma}_1 + \dot{\gamma}_2}{2} \quad (8.57)$$

This set of equation can now be solved to show what happens to the SQUID's IV as the flux through the loop is increased. The figure 8.8 and figure 8.9 shows how the IV curve shifts.

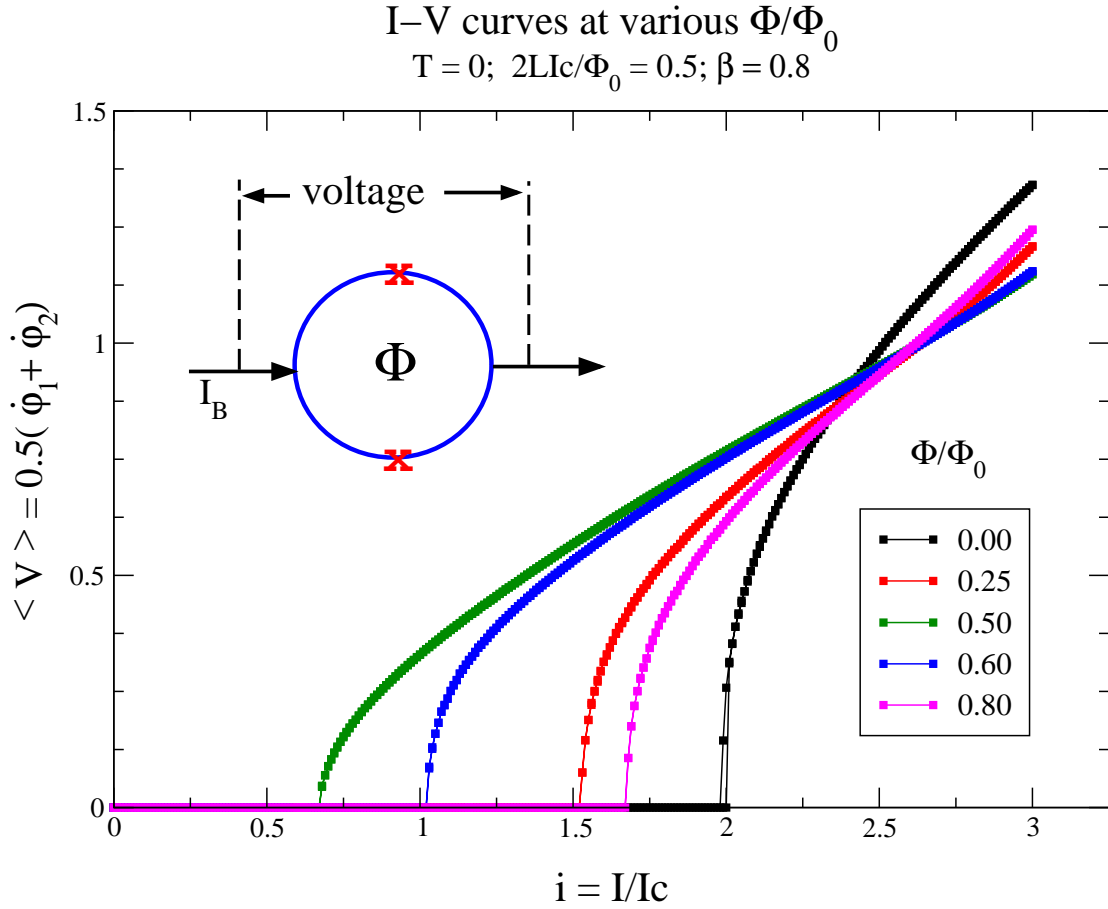


Figure 8.8: Note how the IV curve shifts. In reality most SQUIDs are operated with a current larger than I_c through them. Also the mechanism by which the flux is detected involves a feedback mechanism to keep the flux through the loop constant by changing the current in a separate small coil, that will be explained later.

Thermal noise

We state (without proof) that thermal noise imposes two constraints on the microscopic junction parameters.

- The coupling energy of the junction must be larger than the thermal energy. So

$$\frac{\hbar I_c}{2e} > k_B T \quad (8.58)$$

So at $T = 4.2K$ (liquid Helium) this requires $I_c \gtrsim 1\mu A$ approximately

V- Φ for various self-inductances ($l = 2LI_c/\Phi_0$)

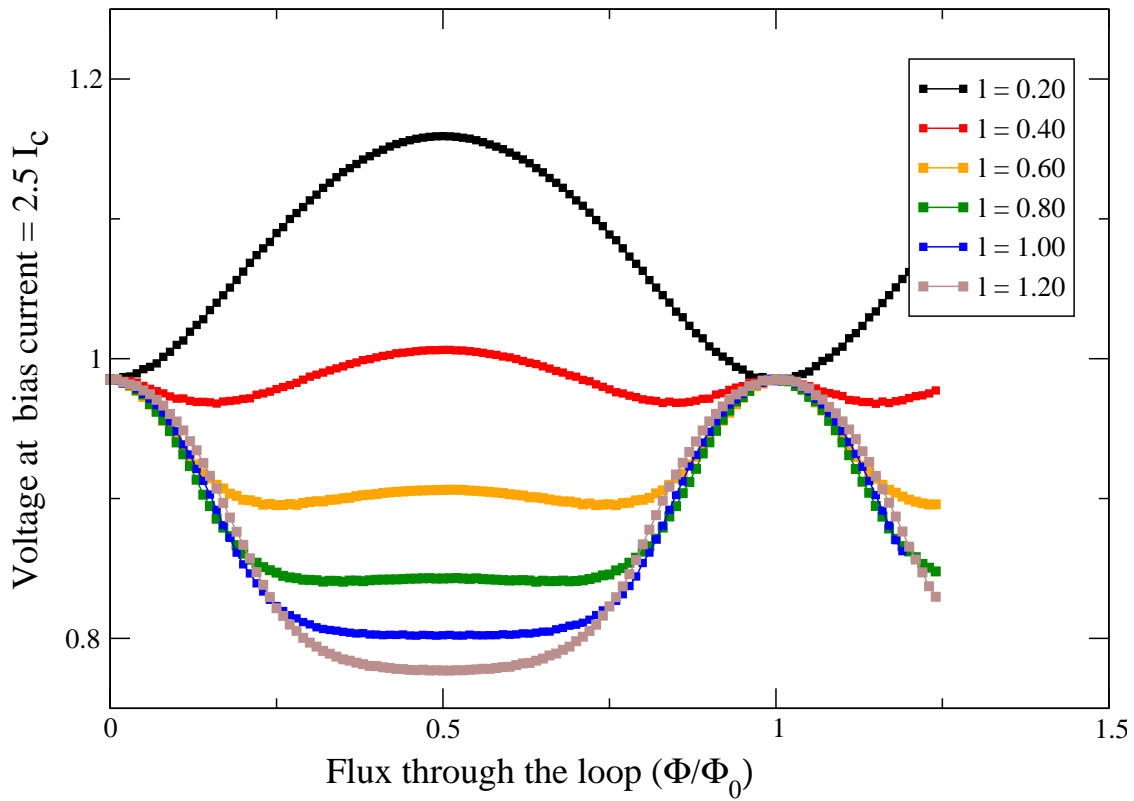


Figure 8.9: See also how the measured voltage would change as the flux is varied.

- The fluctuation in current also implies a fluctuation in the self induced flux. So large self inductance will mean that the fluctuation flux can overrun the effect of the external flux. Simulations show that we need

$$L \lesssim \frac{\Phi_0^2}{5k_B T} \quad (8.59)$$

So at $T = 4.2K$ (liquid Helium) this requires $L < 15\text{nH}$ approximately.

Appendix A

Electromagnetic spectrum

In this chapter we will review the mechanism of generation, detection and importance of electromagnetic waves of various frequencies. We will see that an important boundary exists in an in-between region. Let's first see a very familiar picture:

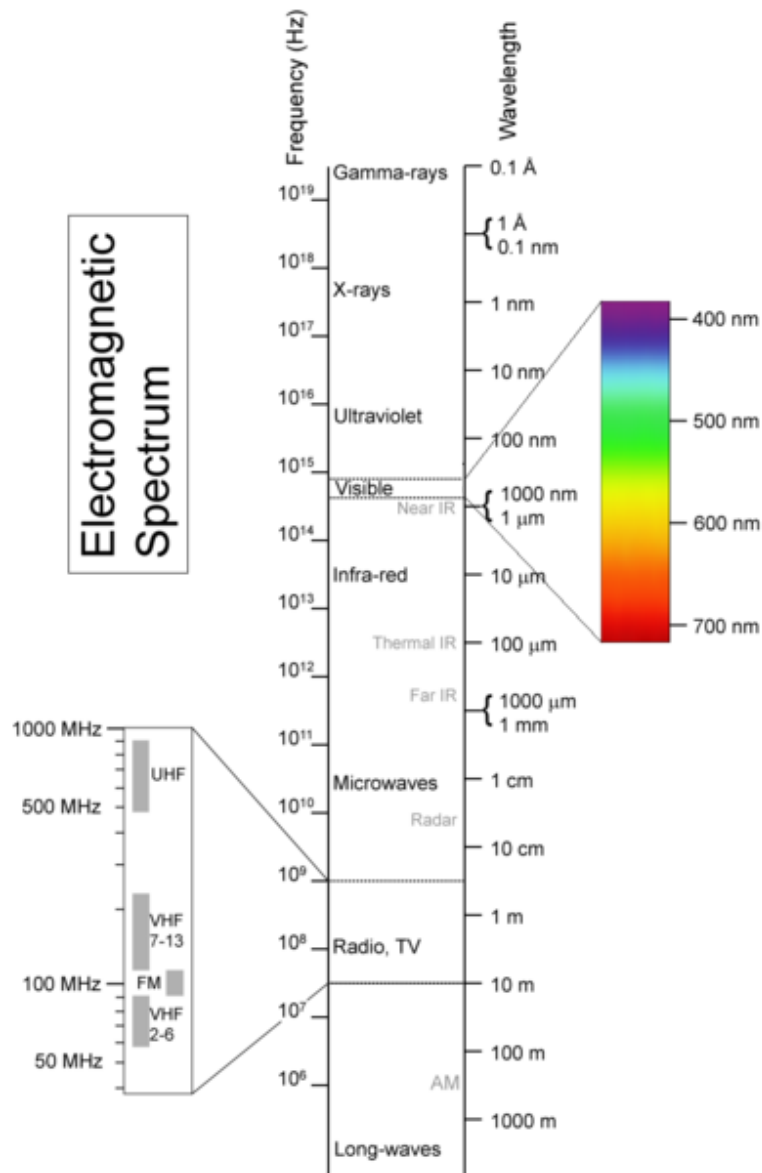


Figure A.1: The electromagnetic spectrum .

Let's start from the bottom of the figure and ask some questions.

- How are very low frequency waves ~ 100 KHz generated? A simple circuit shown would do. At typical audio frequencies and a bit higher this would work fine.

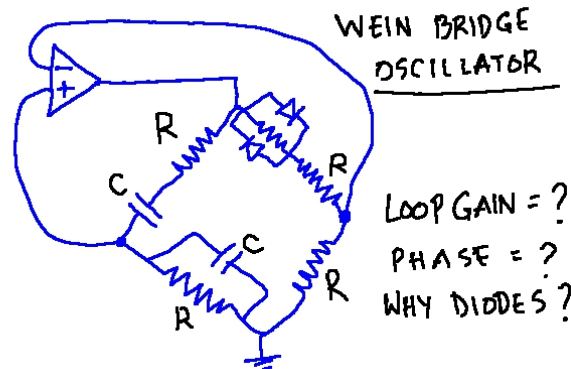


Figure A.2: The simple audio frequency oscillator for few kHz to tens of kHz .

- To go slightly higher one would use LC circuits, why?

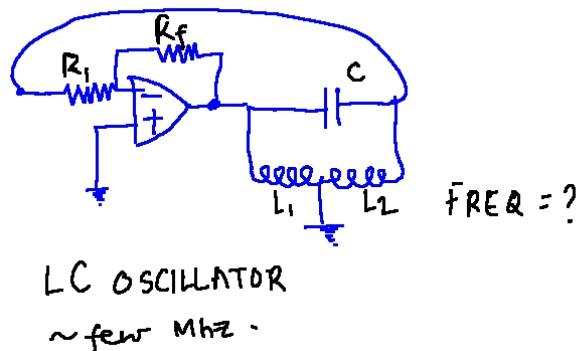


Figure A.3: LC oscillator for few hundreds of kHz - few Mhz .

- As we come to few hundred Mhz the oscillator requires more sophisticated (high speed) transistors for the amplifiers. Why is it so?
- Cell phones work at typically 900 Mhz - 2 GHz range.
- The common microwave oven & wireless routers work around 2.4 GHz.
- The fastest transistors available today approach close to 100 GHz.

Now see the following figure:

- When we want to generate oscillations in the infra-red regime, we rely on fundamentally different mechanisms. Usually this would involve transitions between conduction and valence bands of an semiconductor, or between energy levels of an atom or in some cases just thermal radiation like a tungsten filament bulb. The region of a few THz separates what is conventionally called "electronics" and what is called "photonics".

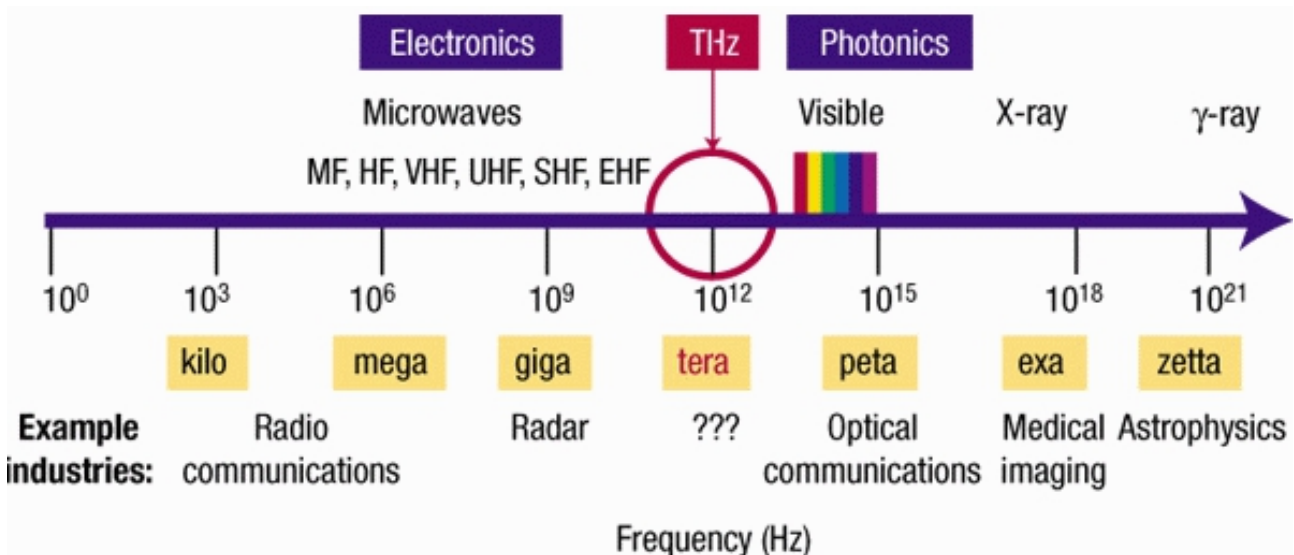


Figure A.4: The terahertz gap. The development of radiators and detectors in this range is a relatively recent phenomena. This is a range where we can neither use electronic oscillators nor atomic/band gap based transitions.

- In the visible range (1-2eV) usually atomic transitions involving outer shell electrons work. Recall the sodium D lines involve outer shell electrons.
- If one keeps going higher and higher in frequencies then one would require transitions from core levels of an atom. For example the $Cu K\alpha$ line is a transition between the innermost and the next higher shell. This is in the X-ray regime ($\lambda = 1.54\text{\AA}$) with energies in the range of 100 keV or so.
- If one keeps going higher one would reach energies characteristic of nuclear transitions - these gamma rays would have energies in the range of 10^6 eV or even more.
- Gamma rays of much higher energies would come from electron-positron annihilations etc.

Finally let us also look at how the atmosphere responds (attenuates) the various frequencies ranges in question.

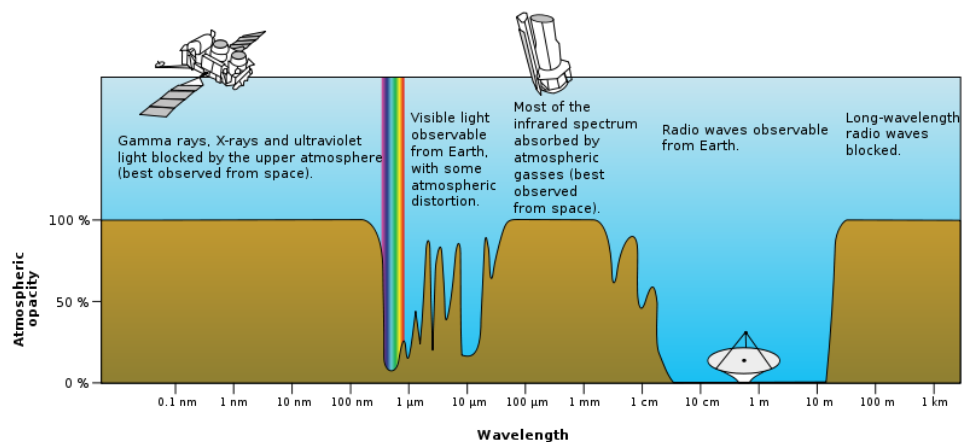


Figure A.5: Which frequencies does the atmosphere allow, and which does it cut off. This often determines the kind of technology we need for communications or to observe radiation in a certain range. The picture is from Wikipedia.

A.1 Mott transition

References:

1. The transition to the metallic state, N.F. Mott, *Philosophical Magazine*, **6**:62, 287-309
 2. The transition to the metallic State, P.P. Edwards and M.J. Sienko, *Accounts of Chemical Research*, **15**, 87-93 (1982)
-

We consider the following thought experiment. Suppose we have a collection of atoms with one loosely bound electron (like s electrons) arranged on a cubic lattice (say) with a very large lattice constant. Now we start shrinking the lattice. One might argue that this can be achieved to some extent by applying pressure to some real material. However, we will make the connection with reality a little later. We want to know whether this material will conduct electric current at $T = 0$. This depends on whether it has free electrons in the conduction band in the limit of $T \rightarrow 0$.

Common sense would say that if these atoms are too far away from each other then the atoms effectively do not see each other and there is no way an electron can move from one atom to another. In a tight binding sense, the overlap integral would be zero. The question is at what separation (and how) does the system switch to being a band metal?

This question has very deep ramifications but let us first try to frame a strategy to solve it.

Band theory alone would not help. Tight binding would just suggest that the bandwidth would go to zero exponentially with distance. If the atoms are very far apart what is the difficulty in conduction? Well for an electron to move from one site to another there would be times when two of these sit on top of one another. This would cost a lot of Coulomb repulsion because the wavefunctions are localised on each site as long as the s electrons are bound to the atoms. But if the binding energy goes to zero then all electrons are delocalised and the question of large repulsion due to confinement in a small volume (one atomic site) does not arise. Recall that we gave a similar argument to rule out double occupancy for donor sites in a semiconductor.

The potential between a lattice site and the outer electron will be the Coulomb potential modified by the screening due to the lattice and the other free electrons already present in the conduction band. This may then be written in the Thomas-Fermi screening approximation as

$$V(r) = -\frac{1}{4\pi\epsilon_0} \frac{e^2}{\kappa r} e^{-q_{TF}r} \quad (\text{A.1})$$

where q_{TF} is the Thomas Fermi wavevector that depends on how many free electrons are already there.

$$q_{TF}^2 = \frac{me^2}{\pi\epsilon_0\kappa\hbar^2} n^{1/3} \quad (\text{A.2})$$

In 3D not every potential has a bound state. There needs to be a minimum "strength" of a potential - before it will develop a bound state. We know that in the limit $q \rightarrow 0$ the screened Coulomb (or Yukawa) potential has bound states - it just becomes the bare Coulomb. On the other hand if q_{TF} is very large the potential drops too fast to develop a bound state. It turns out that the condition for *no bound state* is :

$$q_{TF} > \frac{me^2}{4\pi\epsilon_0\kappa\hbar^2} \quad (\text{A.3})$$

Notice that the combination on the rhs is precisely the inverse of effective (hydrogenic) Bohr radius (a_B). The relation is

$$a_B = \frac{4\pi\epsilon_0\kappa\hbar^2}{me^2} \quad (\text{A.4})$$

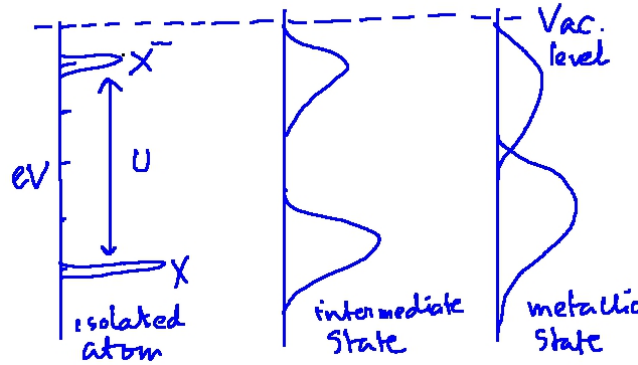


Figure A.6: The Coulomb repulsion energy U when atoms are far away is too large for transport to take place. But the bandwidth increases as the interatomic separation decreases. There comes a point when the overlap integrals are large enough and transport is possible from site to site.

This allows us, to conclude using A.3 and A.2 that if

$$n^{1/3}a_B > 0.25 \quad (\text{A.5})$$

then there is no bound state. Figure A.6 shows another way of thinking of this. In the isolated atom the state with one extra electron lies very high in energy (by an amount U) for reasons given before. Now with this extra input we can calculate the spreading of both the levels, assuming some realistic wavefunctions of both the states (neutral and the one with an extra electron). As the atoms come closer at some point the band width may become sufficiently large so that the two bands overlap. The electron can then move seamlessly from one site to the next. This is the metallic state.

Let us now see the connection of this with dopants in semiconductors. The dopants do not form a regular lattice, but their number can be controlled. For the moment let us forget about order with the following handwaving justification. The bandwidth depends (in a tight binding sense) on the co-ordination number and the nearest neighbour overlap integral. So we have some justification of ignoring what might be happening to the sites far away and just take the average density of sites.

The host semiconductor does the job of keeping the dopants on place and provide a background dielectric constant. The effective Bohr radius and effective mass that we use are the ones appropriate for the host semiconductor.

We can control the number of dopants that we put in. If they are all ionised then the electron density would just be same as the doping density. It is this number that enters eqn. A.5.

An experimental example is shown in Fig. A.7. In general metallic state implies two things:

1. The conductivity tends to a finite value as $T \rightarrow 0$
2. The conductivity increases with decreasing temperature

Microscopically this implies that the electrons are delocalised. Insulating state would mean just the opposite. An important point here is that ultimately temperature plays no role in the problem because we are talking about two possible states (metallic or insulating) in the limit of zero temperature. Thus this kind of phase transition is distinct from the other thermal phase transitions (like water freezing to ice) that we are familiar with.

How well does this hold for doped semiconductors? See the Figure A.8. It agrees remarkably well, inspite of so many assumptions that we made. It tells us that the basic idea of screening by conduction electrons making the Coulomb repulsion cost going to zero, is a very robust one.

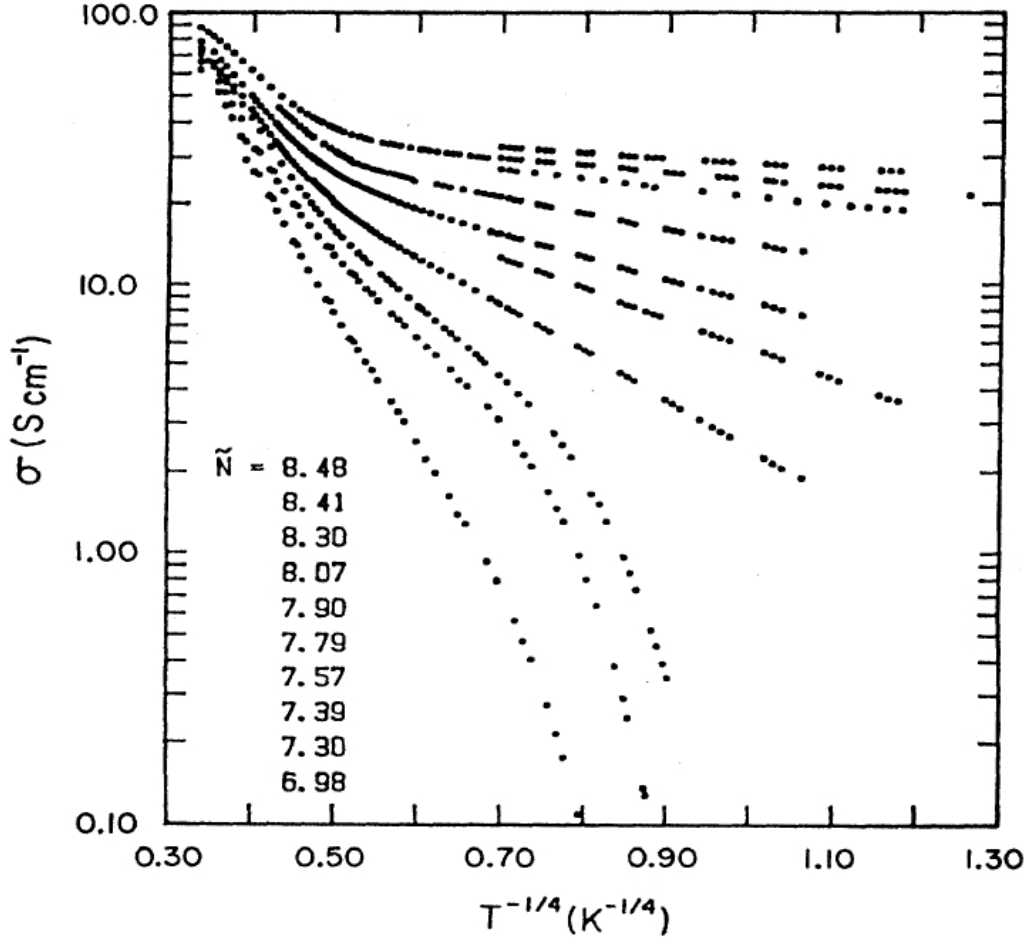


Figure A.7: A plot of $\log \sigma(N, T)$ vs $T^{-1/4}$ for Arsenic donor doped Silicon. Notice how the behaviour changes at higher doping. At low T , $\sigma(T)$ tends to a finite value rather than dropping sharply to zero. The doping levels are in units of 10^{18}cm^{-3} . The data is taken from a paper, "dc conductivity of arsenic doped silicon near the metal insulator transition", by W.N. Shafarman, D.W. Koon and T.G. Castner, *Physical Review B*, **40**, 1216-1231 (1989). This data is on the insulating side, but at the highest doping levels the temperature dependence has almost started flattening off.

These arguments were originally given by Neville Mott starting from late 1940s. There are several review articles that you can get, since a lot of interesting phenomena happens due to the competing effects of electron-electron interaction, disorder and dimensionality.

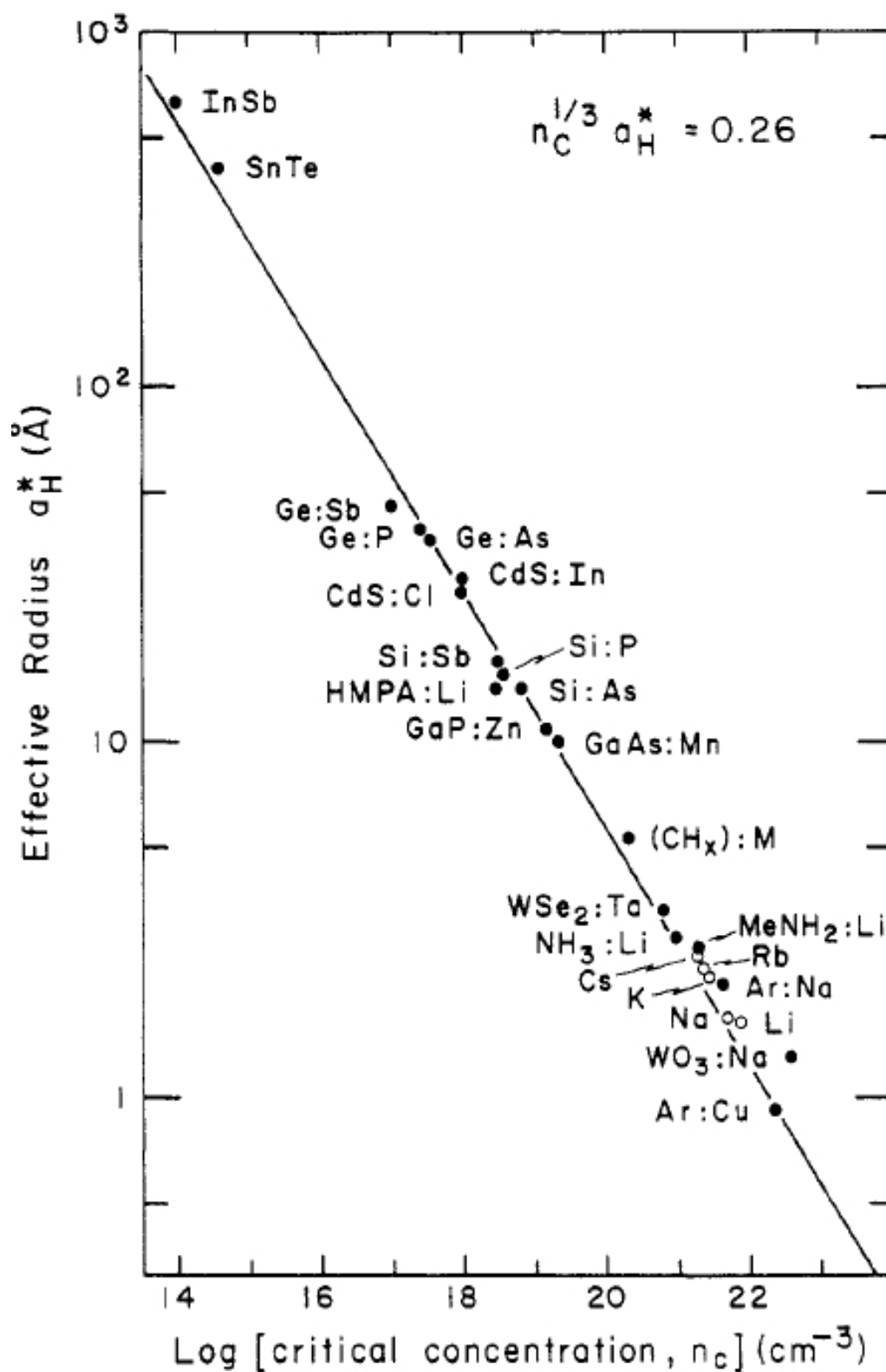


Figure A.8: At what doping density does the transition occur? In this plot the straight line is corresponds to the Mott formula A.5. The data points are experimental numbers at which the transition has been seen to occur. Notice the remarkable agreement for a number of material. The data is taken from a paper "The Transition to the Metallic State", P.P. Edwards and M.J. Sienko, *Accounts of Chemical Research*, **15**, 87-93 (1982).

Appendix B

Many particle systems

The classical view is that one particle, that is part of an N-particle system can be tracked with arbitrary accuracy. Hence they are distinguishable. This is not correct and we need to find a way to write the wavefunction of N particles taking this indistinguishability into account. The fact that particles are either bosons or fermions is reflected in the permutation symmetry of the function of n variables (co-ordinates of the N particles). If we exchange the co-ordinate set that refers to (say) particle 1 with particle 2, the function either picks up a negative sign or stays the same. In general the set of coordinates will consist of the position variable \mathbf{r} and the spin variable σ . Since we will be dealing with electrons only we will denote the spin variables values by $\alpha \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ or $\beta \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. So $\phi_m(\mathbf{r}_5)\beta_5$ will mean that particle number 5 is in a (basis) state ϕ_m and spin state β .

B.1 Determinant and permanent

The wavefunction of a N-particle system needs to be written taking into account the bosonic or fermionic nature of the particles. This is a law of nature and cannot be proved. We can only verify predictions resulting from this.

B.1.1 Fermions

$$\Psi(\mathbf{r}_i\sigma_i) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r}_1)\sigma_\alpha(\xi_1) & \phi_1(\mathbf{r}_2)\sigma_\alpha(\xi_2) & \dots & \phi_1(\mathbf{r}_N)\sigma_\alpha(\xi_N) \\ \phi_2(\mathbf{r}_1)\sigma_\alpha(\xi_1) & \phi_2(\mathbf{r}_2)\sigma_\alpha(\xi_2) & \dots & \phi_2(\mathbf{r}_N)\sigma_\alpha(\xi_N) \\ \dots & \dots & \dots & \dots \\ \phi_N(\mathbf{r}_1)\sigma_\alpha(\xi_1) & \phi_N(\mathbf{r}_2)\sigma_\alpha(\xi_2) & \dots & \phi_N(\mathbf{r}_N)\sigma_\alpha(\xi_N) \end{vmatrix} \quad (\text{B.1})$$

- This expression is called the Slater determinant. If we recall what happens to a determinant, when two rows/columns are interchanged, we can see how the *anti-symmetry* is incorporated.
- We have assumed that the single particle basis states are orthogonal and normalized over the volume of interest.
- Single particle basis wavefunctions may not always be orthogonal, but it simplifies things enormously if they are. Most of the times (for treating electrons in metals etc.) we use the plane wave states $\frac{1}{\sqrt{V}}e^{i\mathbf{k}\cdot\mathbf{r}}$ as the basis, which are orthonormal for different values of \mathbf{k} . But there is one well known example where a non-orthonormal basis is useful - building up the hydrogen molecule from two atomic sites. The (1s) wavefunctions centered on two different atoms are not orthogonal, but serve as useful basis.
- The correct normalisation for orthonormal $\phi_i(\mathbf{r})$ is given by B.1.

- The Pauli exclusion principle is built into this. If two rows or two columns are identical, then the determinant is automatically zero. If we place two particles in exactly the same state - this will happen.
- We also denote the Slater determinant state by the shorthand notation

$$\Psi(\mathbf{r}_i \sigma_i) = \mathcal{A}\{\phi_{i1}, \phi_{i2} \dots \phi_{iN}\} \quad (\text{B.2})$$

where \mathcal{A} denotes the anti-symmetrization operation.

B.1.2 Bosons

We need the completely *symmetrized* product. This is formed by writing out all the possible permutations (P), exactly as if we are expanding the determinant. But we do not include the $(-1)^P$ factor. So all terms occur with a $+$ sign. Such a construct is called the *permanent*, but this name is not very widely used.

- For bosons there is no exclusion principle. We can place as many particles we like in one state.
- Consider a state where we have places n_1 particles in state 1 (denoted by ϕ_1), n_2 particles in state 2 (denoted by ϕ_2)...The correct normalization of this state is given by:

$$\Psi(\mathbf{r}_i \sigma_i) = \frac{1}{\sqrt{N! n_1! n_2! \dots n_Q!}} \left| \begin{array}{c} \vdots \\ \vdots \end{array} \right|_+ \quad (\text{B.3})$$

We leave it an exercise to figure out the combinatorics leading to the normalization factor. The subscript $+$ is a reminder that this is the *permanent*, not a determinant.

- Note the difference in normalization with the fermion case. If there is only one particle in each case, then the normalisation would be identical.
- If all particles are in the same state (say ϕ_3) then the previous expression must reduce to

$$\Psi = \phi_3(\mathbf{r}_1) \phi_3(\mathbf{r}_2) \dots \phi_3(\mathbf{r}_N) \quad (\text{B.4})$$

Show that this follows from the normalization just mentioned.

- We also denote the fully symmetrized state by the shorthand notation

$$\Psi_N(\mathbf{r}_i \sigma_i) = \mathcal{S}\{\phi_{i1}, \phi_{i2} \dots \phi_{iN}\} \quad (\text{B.5})$$

where \mathcal{S} denotes the symmetrization operation.

B.1.3 An important exercise

Let us try to write explicitly the two electron states with free electrons, including the spin functions. from any two spin-orbitals $\mathbf{k}_1 \sigma_1$ and $\mathbf{k}_2 \sigma_2$.

PROBLEM : In a homogeneous system any two body potential is expected to have the form $V(\mathbf{r}_1 - \mathbf{r}_2)$ so the mixing of two determinantal states can occur only if total momentum is conserved. Prove this by taking two initial states with momentum $\mathbf{k}_1, \mathbf{k}_2$ and two final states with momentum $\mathbf{k}'_1, \mathbf{k}'_2$. You need to prove that $V_{\mathbf{k}'_1 \mathbf{k}'_2, \mathbf{k}_1 \mathbf{k}_2} \neq 0$ only if $\mathbf{k}'_1 + \mathbf{k}'_2 = \mathbf{k}_1 + \mathbf{k}_2$.

Now consider the states for which the total momentum $\mathbf{K} = \mathbf{k}'_1 + \mathbf{k}'_2 = \mathbf{k}_1 + \mathbf{k}_2 = 0$. We write the Slater determinants formed with the states \mathbf{k} and $-\mathbf{k}$. The basis states are like

$$\phi_{\mathbf{k}\alpha} = \frac{1}{\sqrt{V}} e^{i\mathbf{k} \cdot \mathbf{r}} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (\text{B.6})$$

So we can have four normalised combinations with $\mathbf{k}, -\mathbf{k}, \alpha$ and β .

$$\psi_1 = \frac{1}{\sqrt{2}} \begin{vmatrix} \phi_{\mathbf{k}}(\mathbf{r}_1)\alpha 1 & \phi_{\mathbf{k}}(\mathbf{r}_2)\alpha 2 \\ \phi_{-\mathbf{k}}(\mathbf{r}_1)\alpha 1 & \phi_{-\mathbf{k}}(\mathbf{r}_2)\alpha 2 \end{vmatrix} = \frac{1}{V} \frac{1}{\sqrt{2}} \left[e^{i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} - e^{-i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} \right] \alpha 1 \alpha 2 \quad (\text{B.7})$$

$$\psi_2 = \frac{1}{\sqrt{2}} \begin{vmatrix} \phi_{\mathbf{k}}(\mathbf{r}_1)\alpha 1 & \phi_{\mathbf{k}}(\mathbf{r}_2)\alpha 2 \\ \phi_{-\mathbf{k}}(\mathbf{r}_1)\beta 1 & \phi_{-\mathbf{k}}(\mathbf{r}_2)\beta 2 \end{vmatrix} = \frac{1}{V} \frac{1}{\sqrt{2}} \left[e^{i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} \alpha 1 \beta 2 - e^{-i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} \beta 1 \alpha 2 \right] \quad (\text{B.8})$$

$$\psi_3 = \frac{1}{\sqrt{2}} \begin{vmatrix} \phi_{\mathbf{k}}(\mathbf{r}_1)\beta 1 & \phi_{\mathbf{k}}(\mathbf{r}_2)\beta 2 \\ \phi_{-\mathbf{k}}(\mathbf{r}_1)\alpha 1 & \phi_{-\mathbf{k}}(\mathbf{r}_2)\alpha 2 \end{vmatrix} = \frac{1}{V} \frac{1}{\sqrt{2}} \left[e^{i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} \beta 1 \alpha 2 - e^{-i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} \alpha 1 \beta 2 \right] \quad (\text{B.9})$$

$$\psi_4 = \frac{1}{\sqrt{2}} \begin{vmatrix} \phi_{\mathbf{k}}(\mathbf{r}_1)\beta 1 & \phi_{\mathbf{k}}(\mathbf{r}_2)\beta 2 \\ \phi_{-\mathbf{k}}(\mathbf{r}_1)\beta 1 & \phi_{-\mathbf{k}}(\mathbf{r}_2)\beta 2 \end{vmatrix} = \frac{1}{V} \frac{1}{\sqrt{2}} \left[e^{i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} - e^{-i\mathbf{k}(\mathbf{r}_1-\mathbf{r}_2)} \right] \beta 1 \beta 2 \quad (\text{B.10})$$

PROBLEM : Which are the singlet states and triplet states? Show that $\psi_2 - \psi_3$ is the only singlet $S = 0$ state. Write down the three possible triplet states. What are the symmetries of the spin and spatial parts of the wavefunctions?

B.1.4 The Hilbert space

Now we have a Hilbert space in which there are $N \times N$ determinantal states with n taking any value. We need a way of moving about between these states. An operator that takes us from an N particle state to a $N + 1$ particle (symmetrized or anti-symmetrized as the case may be) state is called a creation operator. But into which state would the additional particle be added and where shall we place the column corresponding to the new particle. Here is the convention:

Fermions

if

$$\Psi_N = \mathcal{A}\{\phi_{i1}, \phi_{i2} \dots \phi_{iN}\} \quad (\text{B.11})$$

then

$$c_m^\dagger \Psi_N = \mathcal{A}\{\phi_m \phi_{i1}, \phi_{i2} \dots \phi_{iN}\} \quad (\text{B.12})$$

In case ϕ_m is already present in the occupied set $\{\phi_m \phi_{i1}, \phi_{i2} \dots \phi_{iN}\}$, then the new state will vanish.

- Similarly we can add two particles:

$$c_n^\dagger c_m^\dagger \Psi_N = \mathcal{A}\{\phi_n \phi_m \phi_{i1}, \phi_{i2} \dots \phi_{iN}\} \quad (\text{B.13})$$

Since determinants change sign if two columns (or rows) are exchanged, we must have:

$$c_n^\dagger c_m^\dagger \Psi_N = -c_m^\dagger c_n^\dagger \Psi_N \quad (\text{B.14})$$

See figure B.1 for an explicit example with a small number of particles.

This must hold for any Ψ , so we arrive at the *anti-commutation* relation:

$$\{c_n^\dagger, c_m^\dagger\} \equiv c_n^\dagger c_m^\dagger + c_m^\dagger c_n^\dagger = 0 \quad (\text{B.15})$$

- The N particle state itself can be built up by starting with the $N = 0$ particle vacuum state as:

$$\mathcal{A}\{\phi_n\phi_m\phi_{i1}, \phi_{i2}\dots\phi_{iN}\} = c_{i1}^\dagger c_{i2}^\dagger \dots c_{iN}^\dagger |0\rangle \quad (\text{B.16})$$

What is the conjugate of the creation operator c^\dagger ? It is tempting to say that it must be an annihilation operator. But let us see why it is so. Consider two states related by

$$|\Psi_{N+1}\rangle = c^\dagger |\Psi_N\rangle \quad (\text{B.17})$$

hence

$$\langle |\Psi_{N+1}| c^\dagger |\Psi_N\rangle = 1 \quad (\text{B.18})$$

But by definition,

$$\langle |\Psi_N| c |\Psi_{N+1}\rangle = \langle |\Psi_{N+1}| c^\dagger |\Psi_N\rangle^* \quad (\text{B.19})$$

From which the claim follows. It is also straightforward to show that

$$\{c_n, c_m\} \equiv c_n c_m + c_m c_n = 0 \quad (\text{B.20})$$

And

$$\{c_n, c_m^\dagger\} \equiv c_n c_m^\dagger + c_m^\dagger c_n = \delta_{mn} \quad (\text{B.21})$$

When the annihilation operator (say c_m) acts on a Slater determinant state not containing the ϕ_m state, its action is defined to give zero. If the ϕ_m state exists, then it should be moved to the first row of the determinant, unless it is already in the first row, and then removed. This row interchange, if needed, would cost one multiplication by -1 .

The behaviour of the fermionic creation and annihilation operators are now completely defined.

Bosons

For bosons the creation and annihilation operators are expected to do the same job, but the normalization of the boson states are slightly different and the columns in the determinant like expression can be shuffled in any way we like.

$$|\Psi_N\rangle = \mathcal{S}\{\phi_n\phi_m\phi_{i1}, \phi_{i2}\dots\phi_{iN}\} \quad (\text{B.22})$$

for bosons there may be more than one particle in the same state. This is best made clear by writing this expression in a slightly different way (called the occupation number representation)

$$|\Psi_N\rangle = |n_1 \dots n_q \dots\rangle \quad (\text{B.23})$$

Which means that there are n_1 particles in the state labelled 1, n_q particles in the state q . To preserve the correct normalization we need

$$b_q^\dagger |n_1, \dots, n_q, \dots\rangle = \sqrt{n_q + 1} |n_1, \dots, n_q + 1, \dots\rangle \quad (\text{B.24})$$

$$b_q |n_1, \dots, n_q, \dots\rangle = \sqrt{n_q} |n_1, \dots, n_q - 1, \dots\rangle \quad (\text{B.25})$$

The relation between two boson creation and annihilation operators are then commutation rather than anticommutation relation. They are infact very similar to the the ladder operators of the Simple Harmonic oscillator.

Consider states $\phi_\alpha \phi_\beta \phi_\gamma$ and three particles.

1 particle state

$$c_\alpha^\dagger |0\rangle = \phi_\alpha(r_1)$$

2 particle state(s)

$$c_\beta^\dagger c_\alpha^\dagger |0\rangle = \frac{1}{\sqrt{2}!} \begin{vmatrix} \phi_\alpha(r_1) & \phi_\alpha(r_2) \\ \phi_\beta(r_1) & \phi_\beta(r_2) \end{vmatrix}$$

$$c_\alpha^\dagger c_\beta^\dagger |0\rangle = \frac{1}{\sqrt{2}!} \begin{vmatrix} \phi_\beta(r_1) & \phi_\beta(r_2) \\ \phi_\alpha(r_1) & \phi_\alpha(r_2) \end{vmatrix}$$

3 particle state(s) (Not all are listed)

$$c_\gamma^\dagger c_\beta^\dagger c_\alpha^\dagger |0\rangle = \frac{1}{\sqrt{3}!} \begin{vmatrix} \phi_\alpha(r_1) & \phi_\alpha(r_2) & \phi_\alpha(r_3) \\ \phi_\beta(r_1) & \phi_\beta(r_2) & \phi_\beta(r_3) \\ \phi_\gamma(r_1) & \phi_\gamma(r_2) & \phi_\gamma(r_3) \end{vmatrix}$$

$$c_\beta^\dagger c_\gamma^\dagger c_\alpha^\dagger |0\rangle = \frac{1}{\sqrt{3}!} \begin{vmatrix} \phi_\alpha(r_1) & \phi_\alpha(r_2) & \phi_\alpha(r_3) \\ \phi_\gamma(r_1) & \phi_\gamma(r_2) & \phi_\gamma(r_3) \\ \phi_\beta(r_1) & \phi_\beta(r_2) & \phi_\beta(r_3) \end{vmatrix}$$

Notice the row interchange and how each new state is fed to the Slater det from the bottom.

Figure B.1: The correspondence between Slater determinant states and the creation operator actions on the vacuum state $|0\rangle$

B.2 How do all these help?

The hamiltonian for an N-particle system would usually consist of operators that involve the co-ordinates of one particle at a time (like kinetic energy, momentum etc) and some terms like pairwise potential that involve the co-ordinates of two particles at a time. It is possible for three body terms to exist - but they are not very common. We know how the N-particle basis states look like (permanent & determinant), we would need to calculate the matrix elements of the Hamiltonian between such states frequently. It is here that the notation and operators we have introduced come in very handy. Recall that in the harmonic oscillator problem, by introducing the ladder operators, calculation of matrix elements can be made particularly simple - much simpler than integrating and differentiating long combinations of exponentials and Hermite polynomials. In this case (N particle problem) this is so as well.

Consider one electron operators of the type (the total KE would be the simple example)

$$G_1 = \sum_i h(\mathbf{r}_i) \quad (\text{B.26})$$

Let $|\Psi_N\rangle = \mathcal{A}\{\phi_{i1}, \phi_{i2}, \dots, \phi_{iN}\}$ be an N-particle state. We can show that

$$\langle \Psi_N | G_1 | \Psi_N \rangle = \sum_i \langle \phi_i | h | \phi_i \rangle \quad (\text{B.27})$$

using the determinant expansion and counting the terms one by one. But now if we had written out the operator as

$$G_1 = \sum_i h(\mathbf{r}_i) = \sum_{mn} \langle \phi_m | h | \phi_n \rangle c_m^\dagger c_n \quad (\text{B.28})$$

and calculated its matrix elements between the state

$$c_{i1}^\dagger c_{i2}^\dagger \dots c_{iN}^\dagger |0\rangle \quad (\text{B.29})$$

and its conjugate, we would have got exactly the same result. It can be shown that

$$\langle \Psi_N | c_m^\dagger c_n | \Psi_N \rangle = \begin{cases} 1 & \text{if } m = n = \text{occupied orbital} \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.30})$$

In the case where the operator h is just the unit operator (1), we would obtain the operator that counts the number of particles. The number counting operator is thus:

$$N_{op} = \sum_m c_m^\dagger c_m \quad (\text{B.31})$$

With a little bit more, we can show that matrix elements of all two body operators (like interaction energy of a pair of particles) between Slater determinants is correctly reproduced if we rewrite the operator as

$$G_2 = \frac{1}{2} \sum_{i \neq j} V_{ij} \equiv \frac{1}{2} \sum_{klmn} \langle \phi_k(\mathbf{r}_1) \phi_l(\mathbf{r}_2) | V(\mathbf{r}_1 - \mathbf{r}_2) | \phi_m(\mathbf{r}_1) \phi_n(\mathbf{r}_2) \rangle c_k^\dagger c_l^\dagger c_m c_n \quad (\text{B.32})$$

Note the ordering of the operators carefully.